ACKNOWLEDGMENT

I owe my greatest debt to my supervisor,Dr Fitz Taylor, without whom this thesis would have been impossible and Dr M Jaeger,my co-supervisor. I thank Dr M Kingdon and The School Examinations Department; Daud Hassan ; Sharifa Zainab Khalid;Helen Abott; Schools in Colchester; St Helen and St Albans and finally the friends in International Hall, Brunswick Square, who helped me with the proof reading.

ISMAIL SHEIKH SAMATAR

EVALUATING ACHIEVEMENT TESTS ON PROFESSIONAL AND PSYCHOMETRIC CRITERIA

Thesis submitted in fulfilment
of the degree of Doctor of Philosophy

University of London
Institute of Education

February 1983

ABSTRACT

The present study evaluates psychometric and teacher-constructed tests on the bases of separate, but interrelated evidence obtained from:

(a) A factor analysis of teachers' opinions about psychometric vs professional testing and test construction procedures.

(b) Teachers' judgments of test item properties against students' actual performance on the test items.

(c) Empirical evidence obtained from students' performance on tests and ratings of the students on the tests as a whole.

The results are discussed in the light of current controversies.

A problem encountered in comparing psychometric and professional tests was a lack of a third independent criterion for comparison. An attempt has been made to overcome the problem by combining the 3 sets of evidence.

In Part I, the nature of measurement purposes, types, and their interpretations are introduced and discussed in the light of the current opposing views of subjective judgmental and empirical evidential approaches to tests and measurements.

Part II concerns an investigation into teachers' attitudes toward psychometric vs professional techniques of assessment of student performances. The purpose was to obtain teachers' attitudes towards types of measurement techniques, to extract from teachers' responses to testing and test construction procedures, their criteria for evaluating test qualities; and to isolate correlates of teachers' positive attitudes towards tests. One hundred and thirty six teachers took part.

Part III investigates teachers' ability to judge the suitability of items for inclusion into a test. The purpose was

(a) to determine the degree of relationship between teachers' ratings of test item properties and empirical values obtained for the same item properties. The objective criterion against which the adequacy of teachers' judgments of the item properties were evaluated was an actual performance of students on the items. The degree of relationship between teacher ratings and students' actual performance was taken to indicate the competence of teachers to judge test item properties.

(b) to determine the underlying methods used in judging by examining the relative importance given to different item properties. The degree of relationship between each property and the overall rating was taken to indicate the relative importance of each property to the judgment of the overall item quality.

Twenty two teachers and 451 pupils took part in the section.

Part IV investigates the relative worth of psychometric and professional tests as judged by the students themselves. The purpose was to ascertain the amount of improvement as judged by students which arises from statistical refinements of a test. Two psychometric and two professional tests were evaluated on evidence obtained from students' performances and ratings on the tests. Ratings were based on values of validity, fairness and relevance. Seventy six students took part.

The results reveal

(1) the criteria teachers use in judging tests for use with their students

(2) the factors which influence these judgments

(3) teachers' inability to judge psychometric properties

(4) a conflict between the properties teachers and psychometricians value

(5) a lack of any distinct preference by the consumers (pupils) for either method of examination.

These results are discussed in relation to theory and the practice of examining, especially in Somalia.

CONTENTS

INTRODUCTION

THE PROBLEM

The school examinations system in Somalia is heavily dependent on teachers' judgments of student performance through teacher-made tests and through personal observations. The present system of assessment has evolved over many years with historical links with educational procedures in other countries.

In the past, the assessment adhered to employing a selection process whereby classroom teachers subjectively determined the promotion of students from one grade to the next. School teachers provided their own criteria for determining students' mastery of the material taught in the class. The method of assessment is characterized by discrepancies in the educational standards. This led to discontent among parents and the Ministry of Education.

In the late sixties and early seventies, the Ministry introduced common examinations at the end of the 4th, 8th and the 12th grades. However, instead of training teachers for the common examinations, the Ministry commissioned classroom teachers who had no particular expert knowledge of the objectives of the school curriculum or the regional variations particularly of the Somali language which is the medium of instruction. As a result, the system was again heavily criticized. Now, the concern for the school examinations system's objectivity (in the sense objective is defined in this study) is felt both in the classroom and at the top decision-making level.

A major pitfall of the system seems to lie in the way it fails to take advantage of either objective professional assessments or appropriate psychometric techniques.

The purpose of the present research is to review, examine and compare the relative worth of the professional and the psychometric methods of assessing students. The results of the study are intended to improve the current school examinations system in Somalia. The problem for the research will be to determine how much of objective psychometric procedures can be safely left to teachers without loss of examination quality or objectivity.

The study will be done in this country as a training exercise for a much larger study in Somalia.

THE APPROACH

The basic question is refined into three separate manageable questions. Each one accounts for a stage in the research. They are:

(1) Are the objectives of tests the same as seen by professional teachers and psychometricians?

(2) Can professional judgments be used to determine the psychometric properties of a test?

(3) (a) To what extent does each type of test satisfy its objectives as indicated in (i) above?

(b) What is the consumer's (student's) verdict on the perceived

(i) fairness

(ii) validity

(iii) objectivity

(iv) acceptability

of each of these two types of test?

CHAPTER 1

MEASUREMENT AND THE THEORETICAL BASES OF QUANTITY AND QUALITY

One's aim in defining a concept is to clarify it so that a fair amount of agreement on its meaning can be achieved. Clarification or agreement on the meaning of the concept is necessary to facilitate understanding among scientists which, in turn, permits scientific investigation of the concept. Our endeavour to present the reader with some of the most authoritative discussions on measurement pertains to that fact.

The following definitions have been used to clarify the meaning of measurement, as used in education and psychology. Measurement is defined by Stevens (1951) as 'the assignment of numerals to objects or events according to rules'. Measurement is according to Campbell (1938) the assignment of numerals to represent properties of material systems. Russell (1938) sees the issue in broader terms. For him measurement is any method of establishing correspondence between magnitudes and numbers. What is measured in the first definition is an object or an event; property of objects in the second definition and magnitude of property in the third definition.

The latter two definitions are more similar. However, the concept of number (Russell) might be different from that of a numeral (Campbell). For Kerlinger (1973) a numeral may not necessarily convey quantitative meaning while a number does; every number is unique while every numeral is not; there is no exact same number which can be substituted for another number; numerals can always be substituted by other numerals without any change in the empirical meaning represented by the numerals.

Secondly, in Campbell, it is the property of the object that is to be measured. In Russell, on the other hand, it is the magnitude of the property and not the property as such that is to be measured. Also, the terms magnitude and property are not identical, magnitude being an amount of property. The first definition (Stevens) does not mention property nor its magnitude. For him, the simple assignment of numerals to objects according to some rules constitutes measurement, although it may be argued that Campbell and Russell are merely attempting to define the rules for the assignment of numbers and are therefore compatible with Stevens' position. This point has not so far been raised by the various contributors.

Most of the current psychological definitions of the concept of measurement take a viewpoint similar to one of these three (Torgerson, 1958).

Gronlund (1971) for example, defined measurement as quantitative description of behaviour which does not include value judgment. The distinction between quantitative and qualitative descriptions of behaviour can be conceptualized as the following: Person X is a fast reader (value judgment) is a qualitative description of the reading behaviour; on the other hand, X reads 250 words per minute would refer to a quantitative description of the reading behaviour. According to Gronlund, then, qualitative description of characteristics of objects is not a measure.

He distinguished measurement from evaluation. Evaluation, as he defined it, is quantitative plus qualitative description of persons or things. The essential difference between measurement and evaluation, therefore, rests in the subjective input from the observer's stand point. The question will necessarily be raised, however, whether measurement without value judgment is possible.

For Kerlinger (1973) measurement is the assignment of numerals to objects or events according to rules. He defined a numeral as a symbol such as 1,2,3,4,5; I,II,III,IV; a,b,c; good, bad; big, small, and so on. Thus the concept of numeral, in Kerlinger's classification of levels of measurement has broader meaning. It embraces letters as well as numbers, or as matter of fact any one to one correspondence. So when quantitative meaning is assigned to the numeral (symbol) it becomes a number and a valid measurement.

According to this definition, measurement includes nominal scales as well as ordinal, interval and ratio scales. Therefore, Kerlinger disagrees with Gronlund as to what constitutes measurement and agrees with Stevens that nominal scale is a kind of measurement.

At the other extreme, Wood's (1962) definition of measurement is more stringent. It includes only the ratio scale. She defined measurement as some kind of scale along which equal units can be indicated and on which the position of zero corresponds to just nothing of what is being measured. Although Wood made no reference to magnitude her definition is, more or less, in line with that of Russell. The expression 'the position of zero corresponds to nothing of what is measured' implies magnitude.

According to Hemstadter (1964), measurement is a 'process of obtaining a numerical description of the extent to which a person or thing possesses some characteristic'. In his definition, measurement does not include nominal scales.

Nunnally (1970) said 'measurement consists of rules for assigning numbers to objects in such a way as to represent quantities of attributes'. Measurement is 'simple assignment of numbers to data (Horrocks and Schoonover, 1968).

With the exception of Kerlinger and Stevens all the above definitions conceive of measurement as consisting of one or more combinations of ordinal, interval, and ratio scales. Kerlinger and Stevens included all levels of measurement in their definitions. They have suggested that nominal scales satisfy the definition of measurement-assignment of numerals according to rules.

However, the main theme running through most of these definitions is that of thinking quantitatively rather than qualitatively. But there is no obvious contradiction in thinking quantitatively about quality. One needs more than a mere preference to justify the exclusion of qualitative descriptions of objects or properties from the realm of measurement. Among the reasons given for the exclusion of qualitative descriptions are:

1. The property which serves as the basis for classification, need not be interpreted in terms of magnitude (e.g. male, female).
2. Class membership need not be represented by numeral (e.g. male, female, A,B, etc.)
3. Magnitudes of the attributes of scales need not be comparable (e.g. maleness vs femaleness).
4. Naming of objects is arbitrary.

The question to be asked is whether simple description or categorical labels to objects according to their attributes belong to the concept of measurement. The answer to that question will take us back to the purpose of measurement to establish whether that level of measurement achieves its purpose. If the purpose of measurement is to provide information about the attribute(s) of the object of measurement, then qualitative descriptions do provide the desired information about the attributes of objects to be measured which can help the observer arrive at a

meaningful conclusion. Whether one calls that process of obtaining information measurement is another issue. Whatever is decided it is clear that the exclusion of that process leads to the exclusion of an identifiable body of information about the property or event.

Each level of measurement equates measurement with a different aspect of a whole process of obtaining information from objects. The concept of measurement is equated with the process of obtaining scores of measurement, the results of the measurement, the instrument for the measurement or units of the measurement. Here, one needs a more comprehensive statement or definition of measurement which combines all the stages into a total process of acquiring information about the object (attributes, properties, etc.) of measurement.

1. LEVELS OF MEASUREMENT

Whether or not one level of measurement is to be excluded from or included in the definition of measurement does not alter the characteristics of that level of measurement in any way. What might be enlightening for the appreciation of the multiplicity of definitions of the concept, is knowledge of the characteristics abandoned by omitting any level of measurement.

Two important facts underlying the levels of measurement are: that the higher the level of measurement (a) the more statistical methods are applicable, (b) and the more information the level provides. These two facts have a lot to do with the proliferation of definitions and the inclusion or exclusion of any level of measurement.

There are many different ways of classifying measurement. One well-known classification (Stevens) has been described here. For more types

of scales of measurement the reader is referred to Stevens (1951), Torgerson (1958), Kerlinger (1973).

1.1. STEVENS CLASSIFICATION OF SCALES

1.1.1. Nominal Scales

Imagine an experiment on colour blindness to test whether or not certain individuals of a group of candidates are colour blind. The method of scoring colour blindness is to categorize candidates into colour blind and not colour blind. Such dichotomous classification is called nominal measurement because it only names or gives labels. This is the simplest way of perceiving similarities or differences in objects (Guil ford and Fruchter, 1978). The numerals are used only as labels. Words or letters can be used instead (Kerlinger, 1973; Stevens, 1951).

The only statistics permissible at the nominal level are the number of cases and the mode. The data obtained cannot be added or ordered (Stevens, 1951). Most of the qualitative description of objects or properties fall in this level of measurement.

1.1.2. Ordinal Scales

The ordinal scale rank orders the data in such a way that the order of the numbers corresponds to the order of magnitude of the properties (Torgerson, 1958). However, the rank ordering does not necessarily imply equal intervals between values. The numerals with the equal distances usually assigned to objects does not mean that the magnitude of the property of an object is also equally spaced.

For example, if Ali is taller than Abdi and Abdi is taller than Farah, then, Ali is taller than Farah. But the statement does not tell

whether or not Ali is twice as tall as Farah.

Ordinal scales can have a natural zero origin without equal intervals. The numerical zero corresponds to zero amount of the property. In the measurement of attitudes, for example, the zero can be taken as the point where favorableness and unfavorableness diverge.

1.1.3. Equal-Interval Scale

This type of measurement possesses the characteristics of both the nominal and ordinal scales. It also possesses numerically equal distances on interval-scale which represent the equal distances in the property being measured (Kerlinger, 1973).

Most statistical methods are applicable to the equal-interval scale. Statistics permissible at this level of measurement are: mean, standard deviation, order correlations, and product-mo ment correlations (Stevens, 1958).

1.1.4. Ratio Scale

The fourth kind of measurement is the ratio scale. Ratio is an interval with an absolute zero. It is obtained when three things are known, (1) the rank order of persons or objects; the interval between them; and the distance from zero for each person. At that level information about all other levels becomes available.

Stevens was criticized for basing his scale classification on an insufficient evidence. He based his scale identification on certain irreversible mathematical transformation, which does not allow one scale to be classified as ordinal scale on one occasion and as interval scale on another occasion. To Stevens given one-to-one pairing of numbers and

events, scale type is fixed. That is, when values of numbers in a scale are changed, the functions of the scale remain unchanged. For example, if one takes the square root of numbers in an ordinal scale the order is preserved. This invariant principle, according to Stevens is sufficient to determine all the information in the scale.

This proposition was rejected by Prytulak (1975). He cited many occasions on which different scales were employed for the same empirical property. For example, a physicist employs ratio scale to measure the frequency of sound waves but the psychologist employs ordinal scale to measure sound waves;a physicist employs ratio scale to measure electric current and the psychologist employs ordinal scale to measure the electric current; and the same is true of time, temperature etc.

Also Ellis (1966) and Kratz et al. (1971) argued that inadmissible transformations of a scale can still serve the functions and retain the information of the old scale. Therefore, Stevens mathematical transformations to identify scales was ambiguous and unworkable. Prytulak suggested that scales cannot be classified in isolation. He said scale depends on the use to which the events are put.

Stevens was also criticized for giving a misleading impression about the relation of statistics and measurement. He took the position that the type of scale determined the appropriate statistics. This position is also taken by Nunnally (1978). In that view, statistics operate on numbers but is blind to the empirical meaning the numbers represent (Nunnally, 1978; Fraser, 1979). Fraser believes that this situation in which the statistical operation does not consider the empirical meaning affects the conclusions drawn from the results. Fraser did not say how or by how much.

Nunnally (1978), slightly disagrees with Fraser. He said even if misassumptions are made about the scale properties this would have little influence on the results of the scientific experiments. He does not share the doubts expressed by others on whether empirical relational system can be represented by a formal system (scale). Nunnally rejected the assumption of 'real' scale because he said it leads to unanswerable questions. He believes that if the assumptions are good in the first place, violating them would not be harmful. That is, when it is assumed that attitude is measurable on an interval scale and that this assumption is good nevertheless a ratio scale is employed.

2. SUMMARY AND CONCLUSIONS

It is essential to know the meaning of measurement and to promote a common understanding of the concept among scientists as well as others. But at the same time, it is important to know the purposes for which measurement is intended to serve.

In this chapter discussion is confined to the meaning and scope of measurement. One purpose of defining measurement is to develop scales to record and organize human observations. Indeed the development of more organized human observations is believed to be the most important step in the history of measurement.

The second purpose of defining measurement was to promote common understanding of the concept of measurement among scientists, as well as others, to facilitate efficient communication. Common understanding of the concept of measurement among scientists and others facilitates common perception of objects and events. It reduces ambiguity in agreement in observations.

Thirdly, an important aim in defining measurement is to establish an alternative to human observations which, prior to objective measurement, have dominated measurement. The competence of unaided human judgment to fulfil measurement requirements has long been questioned. The quest for empirical measurement is to eliminate human judgment and replace it with more dependable scales. The empirical scales were expected to represent great improvement over human judgment which were seen as fallible and biased. The challenge to human judgment still remains and the quest for more empirical scales continues. But the panacea has been frustrated by limitations encountered in the empirical approach to measurement.

But whether or not the empirical approach represents improvement over human judgment the approach serves another function. It, psychologically, enhances one's confidence in his description of objects and events The application of empirical approach to measurement creates the feeling that one's description of the world is determined by nothing human, but by something upon which subjectivity has no effect. Many important debates in educational measurement are devoted to understanding the relative efficiencies of human judgment vs more empirical approaches.

The theme of the present study is pertinent to the examination of the relative worth of these approaches to educational and psychological measurement. However, before investigating the relative worth of the professional vs psychometric approaches to educational measurement, the types, purposes, values and the limitations of educational tests will be first reviewed for their relevance to this investigation.

CHAPTER 2

THE PURPOSE AND LIMITATIONS OF EDUCATIONAL MEASUREMENT

1. PURPOSES OF EDUCATIONAL TESTING

If everyone's aptitudes, interests and motivation were identical and all individuals were subject to identical environmental influences, there would be no individual differences. Similarly, if individual differences did not exist, testing would never have developed. But the nature of things is such that individuals differ in inherent characteristics and in their exposure to external forces. Tests whether professional or psychometric, are devised to assess these differences.

We test because we are confronted with situations in which decisions have to be made. An employer needs more relevant information about the candidates' desired qualities so that he can decide whom to hire; a classroom teacher needs more information than he personally observes about his pupils so that he can decide whether to proceed to the next unit of instruction or not; and so for every decision maker. In order to make wise and appropriate decisions one must have adequate information to base his decisions on. 'The decision maker who obtains better information before making his decision will get better results' (Cronbach, 1970).

Tests serve the above purpose. By providing some useful information, tests aid decision makers in selecting, classifying or placing individuals; and in predicting future performance. They aid decision makers in evaluating the effectiveness of educational programs and in carrying out research.

If tests serve that purpose by providing useful information for the decision maker, then, their use and subsequent improvements are worthwhile and justifiable. When the knife is appropriate for cutting the piece of meat, sharpening the knife will make it cut better. But

when the knife is basically inappropriate for the job, there is no point in sharpening it.

Now, are tests basically appropriate instruments for gathering adequate information for the decision maker? What is the supporting evidence that tests do provide useful information? If tests gather useful information for decision-making situations,how much better do they do so compared with other methods of assessment? In other words, what is the total contribution of a test in making any decision over all other criteria?

Without trying to give an exhaustive treatment we shall outline some of the ways in which tests are used and found advantageous.

2. TESTS AS LEARNING TEACHING AIDS

Motivating Students

Testing can be useful to students and to teachers in many different ways. One use made of tests in increasing achievement is their effect on student motivation (Ebel, 1975). When students know that they will be tested at the end of an instructional unit they will study more and hence learn more than if they do not know it (Ausubel, 1969). Of course, some approaches to prepare for a test are better than others. But any endeavour to pass a test, however poor it may be, results in better acquisition of knowledge than without it. Even rote recapitulation of facts, the least desirable method of learning, was found to be better than nothing (Wood, 1962). So mere preparation to pass a test is a positive motivating factor. Consequently, frequent testing of pupils must be regarded as a learning aid. Research has shown that motivation is an important variable in academic achievement. It enhances attention,

persistence and meaningful retention of the relevant information (Ausubel, 1969). So for that reason alone, tests are essential components of the learning process.

Several studies support the relationship between frequent testing and student progress. Jones (1923) quizzed an experimental class of students at the end of each lecture. The class frequently tested made scores far higher than the control group.

Turney(1931) and Keys (1934) support the conclusion that with frequent testing better results are obtained. Ross and Henry (1939) tried to determine the relationship between frequency of testing and progress in learning. First, they ensured the comparability of two classes by adjusting previous differences through statistical procedures. The experimental class was subjected to frequent testing (each week). Both classes had a mid-term and a final examination. The authors concluded that in general all frequently tested students achieved better results and in particular students who scored low on a pre-test gained more points than those who scored high on the pre-test.

Fitch, Drucker and Norton (1951) investigated the effect of testing upon motivation of college students to achieve a particular course content. Short weekly quizzes for the purpose of guiding the student's achievement were given in one class. The control class was given only the regular monthly quizzes. In addition to the weekly quizzes, the experimental group had also the regular monthly quizzes.

The results show that the frequently-quizzed students had significantly higher achievement than those quizzed only monthly. In this study the frequently-tested group outperformed the monthly tested group when other factors were accounted for.

The authors interpreted the finding as an evidence that frequent testing motivates students to work hard and hence learn more. Frequent testing also motivated students to attend more discussion groups. The use of frequent testing to increase students' achievements of the subject is also supported by other studies (Standle and Popham, 1960)

Frequent testing was found to be particularly beneficial to less able students (Noll, 1938; Kirkpatrick, 1934; Ross & Henry, 1939).

In one study by Ross and Henry (1939), the ten lowest students on a pre-test gained 68.1 points compared with the 36.1 points gained by the ten highest students on the same pre-test. The two categories of students belonged to the same frequently tested class. In the control class, the gain made by the lowest students and the highest students was much closer together, i.e. 39.3 and 32.8 points respectively. One may interpret the differential gain as being due to a ceiling effect, where the low scoring group of students had more space to move up than the high scoring group.

3. Monitoring Student Progress

Another way of using tests for instructional purposes, is to ensure that pupils are learning. That is, to monitor or make an objective check on student progress. Monitoring student progress leads to two different types of decisions:

(a) The student, from his own relative performance on the test items, discovers his strengths and weaknesses and hence re-studies the content of the topic accordingly. Very often students help instructors by reminding them of the particular areas of a topic

that have not been dealt with sufficiently.

(b) The teacher, for his part, may discover his strengths and weaknesses in conveying the objectives of the lessons to his pupils through testing. On the basis of such information, the teacher feels it necessary to repeat certain parts of the topic; or to use alternative methods of teaching to achieve greater efficiency.

In the teaching-learning process, one further use of tests is to guard against two things listed below which are boring and waste time and energy:

(a) To teach the student what he knows, and

(b) to teach him on a level too far beyond his present knowledge.

The most appropriate method to avoid teaching the student what he already knows or teaching beyond his present knowledge is to test him in advance.

Testing helps the teacher to understand where the next unit of instruction begins, the kinds of remedial treatment needed and whether the method of teaching used was successful. Both standardized and informal classroom tests are useful information-gathering instruments for understanding academic progress. They provide feedback for students, teachers and other persons. Standardized tests are used for broad periodical assessments while classroom tests are used to provide immediate feedback from student progress. They help the classroom teacher to know whether pupils have or have not learned what they have been taught. Only tests which are designed in such a way that they become an integral part of the teaching-learning process can achieve that aim.

4. Diagnoses of Academic Failure

Diagnostic tests are given for the purpose of giving remedial

instruction. Diagnostic tests are for understanding the roots of present difficulty. In the process of diagnosis, the first rudimentary idea about insufficient learning occurs when pupil A is identified as not achieving what his age group is achieving or what he is expected to achieve. But this apparent lag behind one's own group does not automatically indicate that he is having some learning difficulties. All pupils are not expected to have the same achievements even if every one of them is achieving according to his capacity. Someone must be on the lower end of the ability scale. So, it is necessary to distinguish between under-achievement which is caused by real learning difficulties and 'under-achievement' which is due to low intellectual ability. For example, a pupil whose overall attainment is well above the mean of his group may show certain learning difficulties. If he could potentially be at the top of his class, his present achievement of being only above the mean must reflect some kind of learning difficulty.

The method used to differentiate the effects of learning difficulties from effects due to low mental ability is to compare the results of standardized achievement tests of the same person. When his achievement level is significantly lower than his scholastic aptitude level, learning difficulties are thought to exist and further diagnoses to isolate the specific problems are warranted.

So, to simply state that the pupil's overall attainment is not as good as that of his cohort is not enough by itself in diagnostic testing. It is too general and does not tell the decision-maker enough to understand the specific learning problems of the pupils. One's underachievement could be due to many different causes. What is needed is a method to identify the specific deficiencies which contributed to the pupil's underachievement.

What then is the best way of determining these specific deficiencies separately from each other? The way to discover these specific learning difficulties is to adminster a battery of tests to the subjects to be diagnosed. Each test is constructed in such a way that it gives evidence on the strengths and weaknesses of a basic skill of the pupils. As it contains more items representing each skill to be measured, a battery of tests has better chances of locating specific deficiencies (Thorndike and Hagen, 1977).

Analysis of profiles based on achievement tests can also locate areas of weakness. Similarly, criterion-referenced tests can be utilized for diagnostic purposes to provide information about an individual's mastery or non-mastery of particular skills (Anastasi, 1976). The third step in the diagnostic process is to identify the causes of the learning difficulties. It is the aim of diagnostic testing to remove the roots of the learning difficulties. Difficulties in learning may be attributable to many different factors. They could be attributable to physical handicaps, lack of motivation, improper teaching, disadvantaged home background, etc. The nature of the learning difficulty found may provide the first clue to the root cause. The root cause may turn out to be the pupil's ill health or other physical handicap; low scholastic aptitude, language skills, lack of adequate motivation, some emotional maladjustments, inadequate working habits, or exposure to a variety of environmental factors.

The remedial treatment to be prescribed will always depend on the nature of the specific learning difficulty found. From there, the effect of remedial treatments chosen can always be monitored by further diagnostic testing. Further diagnostic testing provides evidence on the correctness of the treatments. On the basis of the kind of deficiency revealed in the treatment, further decisions could be made about the original treatment.

5. TESTS AND CURRICULUM REFINEMENT

The function of tests in decision making is also found in the area of curriculum evaluation where tests are aids to evaluate curricula and to monitor their effectiveness. Through testing, teachers assess the attainments of pupils as well as the effectiveness of the methods of teaching. Then, on the basis of the information provided by the tests, teachers may feel it necessary to modify their methods. Similarly, the content and structure of the curriculum may be changed according to the learning outcome. For any treatment, whether it is a school curriculum or a research experiment, tests give more dependable information for its evaluation than subjective judgments (Cronbach, 1970).

A good example, where tests are used as aids to evaluate the effectiveness of a whole education system is in the case of the National Assessment of Educational Progress (U.S.A.). The aim of this assessment program is to gather information on the knowledge, skills and attitudes of American students. The testers thought that the best way to obtain relevant information to assess whether students were prepared in the basic subjects, was to give a battery of tests each representing areas mostly taught in schools. The same battery of tests are repeated periodically to monitor whether performances observed earlier have improved or regressed. In other words, tests are utilized to give evidence on the present knowledge and skills of students and how this knowledge and/skills change over time these (Brown, 1970).

A second example where tests are used to aid the evaluation system is in the case of the Assessment of Performance Unit (Britain) which is designed to assess the overall standards achieved by students in different parts of the curriculum (Broadfoot, 1979).

A project of still larger scope, where tests are used to aid evaluation, is the IEA's testing program to evaluate the relative efficiencies of education systems of 20 different countries. The aim of the International Association for the Evaluation of Educational Achievement's programme is 'to develop standard measures of educational achievement, describe the achievement level in various countries, and identify the factors that account for cross-national differences in achievement' (Brown, 1970).

In all the studies (NA, APU, IAEEA) it has been indicated that useful information can best be obtained from testing, which can be used to improve education. In general, tests are preferred to other methods as the best instruments which provide the most dependable information for the evaluation processes.

6. PLACEMENT

Tests aid decision makers to infer the category to which an individual will belong, so that individuals can be grouped according to abilities or to other traits of interest. In schools, students are classified and placed in order to maximize learning. On the basis of the information provided by the tests, pupils may be grouped in different streams, some may be placed in remedial instruction programmes. Or students may be grouped in accordance with the talents they have shown in different areas of the school curriculum.

In jobs, candidates are classified and placed in such a way that their talents are maximally utilized. In counselling and guidance, the same information facilitates appropriate decisions made about advisees by their advisors, or individual advisees about themselves.

In all other situations related to placement decisions, tests give

similar information to help the decision maker to arrive at appropriate placements of individuals in their respective positions.

7. OTHER USES OF TESTS

Tests are also used by the individual to make decisions about himself; to achieve social benefits; to make efficient use of resources; to communicate student progress to parents and the public and countless other situations.

Every individual makes decisions about himself concerning his choices and plans. To make wise decisions about oneself one must know what type of person one is; what abilities, interests and temperament one has. He must also know how well his characteristics match the many options of life open to him. Tests help the person to know about himself by providing valuable information on which he can base decisions about himself.

8. Social Benefits of Testing

The social benefit of testing is to protect positions of social importance from persons who are incompetent. To place individuals in occupations which they cannot cope with, would bring harmful consequences to society. Imagine the disaster that would befall a society which has incompetent medical practitioners or incompetent airline pilots etc.

To avoid malfunctioning, society demands guarantees against individuals who lack the knowledge and the skills to execute the tasks they have to undertake. To ensure competence and to safeguard against ineptitude, tests are used to assess and certify the knowledge and skills of candidates.

9.1. Efficient Use of Social Resources

To try to train individuals for skills they are unable to acquire is a wastage of resources which society tries to avoid. So training must be given to persons who will profit from it. Such persons are selected for the training programmes through testing the candidates.

9.2. Accountability

Nations invest a great deal of money in education. So, the education system is accountable for the money spent to produce desirable standards (Broadfoot, 1979). Members of any community are concerned with the education of their children. They like to know whether children are learning what society expects them to learn. They demand evidence of student progress. Persons who are accountable for the education of the youth communicate student progress to the public through test results. So tests are means of settling disputes between parents, or the public, and the educational system.

10. LIMITATIONS AND ABUSES OF TESTS

Critics of tests point to some limitations and abuses of tests. Criticisms directed against tests are many and we do not intend to give an exhaustive treatment of all the points raised. We shall only mention a few of the more publicised limitations and abuses of tests.

10.1. LIMITATIONS

10.1.1. Tests Measure Trival Knowledge

In the schools, tests are criticised as measuring only factual knowledge rather than the more important educational outcomes, such as

comprehension, problem, originality, creativeness etc. Tests are said to deny creative persons the opportunity to demonstrate their creativity (Hoffman, 1962). The result is that geniuses would never be identified. Also, other desirable human traits such as honesty, co-operativeness, compassion and sensitivity which schools have to inculcated in students, if they are to become valuable members of the society, cannot be measured by tests.

10.1.2. Tests become an end in themselves

In learning, tests become an end in themselves. Pupils and teachers direct their efforts and interest towards passing tests rather than intrinsically understanding the subject (Ausubel & Robinson, 1969). Tests encourage competitiveness and excessive external motivation. They curb independent thinking and encourage conformity. They encourage mechanistic decisions, rather than rational decisions to be made about pupils, and the inaccuracies of the decision-makers to be obscured and overlooked (Ebel, 1975; Brown, 1970).

10.2. ABUSES

10.2.1. Tests put people's destinies in the hands of a few test experts

Tests place the destinies of many people in the hands of a few test experts. Information provided by tests can only be interpreted by test experts. Ordinary persons are in no position to understand how tests are constructed, standardized or validated. As a result of their monopoly of knowledge about tests, test experts become so influential that theyvirtually control school curricula and hence the future of every student (Ebel, 1975; Barclay, 1968; Dubois, 1964).

10.2.2. Attach undesirable labels to pupils

Tests jeopardize the future social status of some students by attaching a label to their intellectual capacity while what the tests measure may be only learned skills and not potentiality (Goslin, 1968). The belief that a person's IQ does not change reinforces predeterministic and rigid classification of persons. For example, a child who as the result of his high score on a test was given special attention by parents or teachers or enrolled in a better institution, etc. would consequently perform better than that who received a lower score at the beginning and was given worse treatment. When an inferior label is attached to him, the pupil's self-esteem and educational motivation is likely to be badly harmed.

10.2.3. Tests undesirably discriminate between groups and individuals

The most serious limitation of tests is the fact that most of the tests are culturally (and in other ways) biased and unfairly discriminate between peoples of different cultures, socio-economic backgrounds, sex etc. (Goslin, 1968). As tests are constructed by members of the dominant culture, test results reflect the values and the skills of that culture (Broadfoot, 1979). Therefore, most tests cannot measure the aptitudes and the skills of the minority groups (Jensen, 1980).

11. Other criticisms

Other criticisms levelled against tests include that test items are often ambiguous and trivial (Hoffman, 1962); that some traits are difficult to define and hence impossible to be measured by tests; that test results are contaminated by extraneous factors, such as race, level of motivation, pupils' attitudes etc; that tests confirm teachers'

expectations of students and reinforce biases; and that tests of personality invade privacy by allowing others to have access to personal information on the individual (Dunnete, 1963).

12. SUMMARY

This chapter concentrated on answering questions asked at the beginning of the chapter. That is, whether tests are basically appropriate instruments for measuring certain human traits. The literature reviewed has revealed that tests provide more useful information than any other technique of assessment. The review supported the view that tests are the most preferred information-gathering instruments for most evaluation programmes.

In schools, for example, tests are more suited than any other assessment technique to monitoring academic progress; to exposing the nature of strengths and weaknesses of an instructional programme; and to allowing instructions to take place at appropriate levels; tests can be so easily manipulated to probe many areas of concern without necessarily examining each area at a time; through tests, decision-makers are able to make more precise groupings of individuals according to particular attributes in question.

It is only through tests that large scale evaluation of curriculum effectiveness can be made; that an efficient use of resources can be achieved; that society can best protect positions of social importance from incompetent persons; and that questions of educational accountability can be settled between education authorities and the public.

However, tests are not without criticisms. Tests have been criticised for their failure to measure all the desirable traits to be

measured; that tests measure the most trivial qualities of man; that tests become an end in themselves whereby students and teachers are totally engaged in strategies to pass tests rather than in an acquisition of meaningful knowledge; that tests curb independent thinking and encourage conformity.

Tests put undesirable labels on persons which predetermine their future; tests also put man's destiny in the hands of a few test experts; and tests undesirably discriminate between groups and individuals.

Despite these criticisms, tests are the most appropriate techniques of measurement. No-one has yet found a better substitute for tests. The present study will then address itself to the evaluation of the various alternative techniques available within the domain of tests.

CHAPTER 3

TYPES OF EDUCATIONAL MEASUREMENT

1. APTITUDE AND ACHIEVEMENT TESTS

Achievement tests measure what has been learned or the mastery of school subjects. On the other hand, aptitude or intelligence tests measure potential (Brown, 1970).

But the distinction between achievement and aptitude or intelligence tests is not as simple and clear cut as stated above (Womer, 1975; Glaser, 1962; Brown, 1970; Jensen, 1980). The two tests have many elements in common. First, both aptitude or intelligence and achievement tests measure performance. The argument is this: whatever the person is able to perform must have been acquired by him sometime in the past. For example, to be able to give correct responses to questions from the vocabulary or information sections of an intelligence test, the person must have learned the words, the names of people and places which are asked for. One cannot expect the testee to respond correctly to concepts which he had never been exposed to directly or indirectly.

Secondly, there is and should be a considerable correlation of about .50, on the average between intelligence and achievement test scores. This relationship between the two types of tests indicates that they are directly or indirectly measuring the same thing whatever that thing might be.

Thirdly, the two tests can sometimes be used interchangeably. Achievement tests can provide reliable evidence on the individual's intelligence (Jensen, 1980). On the other hand, we know that intelligence tests predict future achievement.

Then, how can one differentiate achievement tests from intelligence or aptitude tests? In spite of a great deal of overlap of the function of intelligence and achievement tests there are still essential ways in which achievement tests differ from intelligence tests.

Jensen (1980), contended that the two types of tests are distinguishable. He outlined the points of distinction between achievement and intelligence tests as follows:

"1. Intelligence or aptitude tests have items heterogeneous and broader than achievement tests, which have items more specific and usually confined to specific types of skills and knowledge associated with formal schooling.

2. Intelligence tests sample cumulated knowledge and skills from the individual's past experience, whereas achievement tests sample knowledge acquired in the recent past.

3. Intelligence and aptitude tests predict future intellectual achievements, even though the contents of the achievement have nothing in common with the aptitude tests.

4. Most intelligence measures are more stable across time and are less susceptible to the influence of instruction or training than most achievement measures."

1.1. ACHIEVEMENT TESTS

1.1.1. Objective Tests

Achievement tests may be classified as objective tests, essay tests, oral examinations and performance tests or work samples. Each technique has its own advantages and disadvantages, some of which will

be discussed later.

In objective type tests, the questions are presented in such a way that the correct answer is predetermined. The testee responds to the questions by recalling or recognising the correct answer.

1.1.1.1. Advantages of Objective Tests

Objective tests are now very common because they have several advantages over other methods of assessment. First, objective tests permit more adequate representative sampling of the content to be covered. Second, they eliminate subjectivity and variability in scoring. They allow an invariable criterion for scoring to be made available in advance. Thirdly, more questions can be given to and answered by the testee in a relatively much shorter time while the testee still has more time to think. Fourthly, scoring is easy and saves the examiner time for other things. Fifthly, objective test items lend themselves to item analysis so that the difficulty, discrimination, reliability, validity etc. of the test can be improved for future use (Ausubel, 1969; Child, 1973; Gronlund, 1971; Nunnally, 1970).

Other advantages claimed for objective tests are:
that objective tests are more valid and more reliable than other methods of assessment; that more expert item writers are available; that more objective tests are available from commercial firms (Nunnally, 1970); that more objective items can be retrieved from item banks; that, contrary to criticisms, objective tests can measure almost any performance accurately etc.

1.1.1.2. Disadvantages of Objective Tests

The disadvantages of tests have been briefly discussed in Chapter Two. However, we shall repeat some of these criticisms, particularly those aimed at objective tests. Some of the limitations of objective tests include their inability to allow the examinee to organize ideas and to solve problems; that objective tests are adapted to measure only verbal learning outcomes; that they are ambiguous and often prevent the examinee from explaining his choices of answers; that they do not test the examinee's ability to organise his ideas or to present an argument.

1.1.1.3. Standardized Tests

Objective tests are divided into standardized (psychometric tests) and teacher-made tests (professional tests). The former mainly differ from teacher-made tests in that they are intended to be used over a period of many years and to cover a broader range of skills and understandings (Gronlund, 1971). Therefore, test publishers take a great deal of time to develop standardized tests. Standardized tests sample wider educational objectives common to many schools. They provide normative data that permit comparisons of scores across schools and individuals. So, standardized achievement tests are more appropriate in areas which involve some sort of comparison such as guidance and counselling, selection or curricular decisions (Thorndike and Hagen, 1977; Horrock and Schoonover, 1968).

1.1.1.4. Teacher-Made Tests

Classroom teachers like to assess the academic progress of their pupils. When they try, teachers usually find standardized tests too broad

to assess local instructional programmes or suitable standardized tests too expensive to obtain. Teacher-made tests are designed to fit these specific situations (Barclay, 1968). Unlike standardized tests, classroom tests are not designed to give broad comparisons across schools, (Horrocks and Schoonover, 1968). They are only adapted to evaluate contents unique to a particular classroom.

1.2. INTERPRETATION OF ACHIEVEMENT TESTS

For the rest of this study we shall be more concerned with the achievement tests.

1.2.1. PERFORMANCE STANDARDS

Test scores, as they stand by themselves have no meaning (Travers, 1955; Thorndike and Hagen, 1955). A score of 100 standing by itself could mean many different things ranging from perfection to failure. The meaning of a test score then depends on what other information is available, that is to say on the other scores obtained in the same test, or on how the score is compared with an agreed level of mastery set in advance. In other words, test scores acquire meaning only when they are compared with each other or with an absolute standard criterion (Thorndike and Hagen, 1977). This additional information is necessary for the interpretation of all tests (Ebel, 1962).

These two approaches to interpreting achievement test scores are called norm-referenced and criterion-referenced interpretation of test scores, (Glaser, 1962). In the criterion-referenced approach, the degree to which the testee has attained criterion performance is the point of reference, or an absolute standard set in advance. On the other hand, the information provided by the norm-referenced approach is the relative ordering of the testees. The two approaches to test

interpretation, with new developments in the criterion-referenced tests, are discussed in the next chapter.

1.2.1.1. Essay Tests

Stalnaker (1951) defined the essay examination question as a 'test item which requires a response composed by the examinee'. A similar definition was given by Ebel (1971). He said 'an essay test presents one or more questions.....that require extended written responses from the person being tested'.

1.2.1.2. The Main Differences between Essay Tests and Objective Tests

Essay tests' main difference from objective tests is that the examinee supplies the answer to the essay examination question. In other words, the degree of freedom the examinee has to answer the essay examination question is the distinguishing feature of essay tests. Questions included in the essay examinations usually require the examinee to write extended answers. However, some questions might be characterized by limited responses.

Essay testing is a technique which has a significant place in educational assessment. Like other measurement techniques, essay examination questions are given to pupils with the intention to elicit some information about the examinee's behaviour. Essay questions are used to measure learning outcomes as the result of educational experiences. They are particularly useful in testing students' ability to organise ideas and to evaluate them critically. They permit original and independent thinking (Hoffmann, 1962). They allow the examinee to express himself so as to convince the reader (examiner).

Preparation for essay tests has positive influences on study habits. Students concentrate on the main ideas and trends rather than on the details in the subject matter that would be covered by the essay test. They try to understand underlying relationships and attempt to draw logical conclusions (Coffmann, 1971).

There are other reasons why essay testing continues to be popular in educational measurement. One reason could be attributable to the long history and the well established tradition of the use of essay tests. Secondly, essay tests are relatively easy to set (Mehrens, 1975; Ebel, 1979).

It is true that most teachers know that scoring essay examination questions is very tedious. However, many teachers are tempted not to forgo present convenience (easy preparation now) for one in the distant future. They ignore the future pain of scoring essay papers. Thirdly, the test constructor is said to be relatively more secure in preparing essay questions. That is, the deficiency of the essay question is less observable than that of the objective question. Fourthly, since the points assigned to each question are personally decided by the scorer himself, as he proceeds through the essay papers, the examiner can manipulate the distribution of the test scores. Such manipulation often saves examiners from the embarassment of their tests being too difficult or too easy for the students (Ebel, 1979). Fifthly, many teachers may not be aware of the limitations of essay tests at all.

As for their disadvantages, essay tests are less adapted, than objective tests, to measure knowledge of concepts and information (Gronlund, 1971). Since only a few essay questions can be asked and answered by the examinee in any one session, essay questions do not sample the domain comprehensively. This restriction of the number of essay questions that

could be administered to students at any one session brings an element of luck into the examination results. As a consequence, students' performance becomes dependent on the examinees' luck in the particular questions asked rather than on their true abilities. This situation where some students do better on certain questions than on others would improve as the number of questions increases. Inclusion of more short answer essay questions is another way of overcoming the disadvantages of the limited sampling.

Scoring essay type questions is very laborious and time consuming. At the same time, the validity and the reliability of the scores are usually unsatisfactory (Gronlund, 1971). Low reliability of the scores is the most serious disadvantage of essay tests. If better reliability of the scores is demanded the cost of scoring the essay questions gets higher (Mehrens, 1975).

Essay tests reward only those students who write neatly, fluently and persuasively enough to influence test markers' judgments in their favour. On the other hand, lack of neatness, improper punctuation, bad handwriting, inappropriate paragraphing and the like secure examiners' bias against papers so characterised (Marshall & Powers, 1969; Ebel, 1979).

Marshall (1967) tried to determine the influence of errors of grammar, spelling and punctuation on essay examination. He asked 700 teachers to grade thirteen essay forms all identical but differing in the types and the number of the composition errors mentioned above. The teachers were given explicit directions to concentrate only on the content of the essay compositions when assigning grades to the individual essays. He found that any error inserted into an essay had the effect of lowering the grade assigned to that essay. However, the effect of these errors were not multiplicative. Results indicated that teachers' judgments were

influenced by these errors even when teachers attempted to grade the content alone.

The variations observed between scorers, or in the same scorer from time to time, are due to the tendency of different examination markers to concentrate on different aspects of the examination paper to be graded; or raters differing in severity of grading; or raters' scores being influenced by the order in which they read the individual essays etc. Similarly, single raters have the tendency to assign different scores on different occasions. These variations increase as the essay question permits greater freedom of response.

Hulter (1925) investigated whether teachers are consistent in grading essays. He used five English essay compositions which had already been evaluated and standardised (Hudelson, 1925). These essay compositions were sent to 30 English teachers of 7 years teaching experience, on the average, to grade. The compositions but now rearranged were sent to be graded by the same teachers, on a second occasion.

Hulter found that teachers were not consistent in grading the papers. He concluded that teachers' gradings were guesses, some of which could be good and others bad. He suggested that essay tests should not be used for promotion. For promotion purposes, other instruments such as objective tests should be used instead. Since different teachers concentrate on different aspects of the essay paper, he recommended that teachers should be given common aspects of the test paper to be graded and the weights that should be assigned to each section of the essay paper.

In summary, the large variations and the low reliability observed in essay test scores are mainly due to limited sampling of the content covered by the test; the indefiniteness of the tasks set by the essay questions; and the subjective scoring of essay answers (Gronlund, 1971;

Stalnaker, 1951; Coffman, 1971).

1.2.1.4. Performance Tests

Another group of achievement tests are the performance tests. A performance test has been defined as a 'sequence of activities aimed at modifying the environment in specified ways' (Fitzpatrick & Morris, 1971). These are tests with which criterion situations are simulated. Performance tests have been used to evaluate skills related to areas which do not lend themselves to be easily measured by pen-and-paper tests. These areas include industrial arts, vehicle operations, art (drawing and painting), music, sport and many others (Fitzpatrick and Morris, 1971) that pose similar measurement problems.

The purpose of performance tests is to assess certain educational outcomes where other methods of assessment have failed, and to simulate real life conditions of performance. The value of performance tests lies in their ability to simulate comprehensively all aspects of a real life performance of a person, expected to have had the necessary training and experience, in executing a specific task, and to represent these aspects of the real life performance faithfully. In a good performance test, the testee gives most of the desired responses which represent the criterion task.

2. SUMMARY

As discussed in this chapter, there are several types of achievement test the relative worth of which one can always investigate. Essay type tests can be compared with objective type tests; oral type tests with the essay or the objective type tests, etc. Tests can also be classified as professional and psychometric tests.

The present study addresses itself to the investigation of the latter classification. The study compares teachers' attitudes towards psychometric vs professional approaches to tests, testing and test constructions. It correlates teachers' subjective judgments with students' actual performance on an objective test. The two specific types of tests actually compared are professionally constructed multiple choice tests and psychometrically constructed multiple choice tests.

Then, the relative worths of the professional vs psychometric approaches to tests, testing and test constructions will be evaluated on the evidence obtained.

CHAPTER 4

INTERPRETATION OF ACHIEVEMENT TESTS: NORM-REFERENCED AND CRITERION-REFERENCED MEASUREMENT

1. INTRODUCTION

The debate on norm-referenced and criterion-referenced measurements is another effort to refine educational measurement. The merits of the two approaches are discussed within the frame of the controversy between psychometric vs professional methods of educational measurement. Among those who prefer criterion-referenced measurements to norm-referenced measurements are those who advocate empirical procedures and those who would put more confidence in human judgment. The aim of this chapter is to examine the relative advantages claimed for psychometric vs professional methods in the context of criterion-reference testing.

2. PERFORMANCE STANDARDS

Test scores, as they stand by themselves, have no meaning (Travers, 1955; Thorndike and Hagen, 1955). A score of 100 standing by itself could mean many different things, ranging from perfect to failure. The meaning of a test score then depends on what other information is available. That is, to say on the other scores obtained in the same test, or on how the score is compared with an agreed level of mastery set in advance. In other words, test scores acquire meaning only when they are compared with each other or with an absolute standard criterion (Thorndike and Hagen, 1977). This additional information is necessary for the interpretation of all tests (Ebel, 1962).

These two approaches to interpreting achievement test scores are called norm-referenced and criterion-referenced interpretation of test

scores (Glaser, 1962).

3. DISTINCTION BETWEEN NORM-REFERENCED AND CRITERION-REFERENCED MEASUREMENTS

In criterion-referenced approach, the degree to which the testee has attained criterion performance is the point of reference. Performance in relation to perfection is the point of reference.

On the other hand, the information provided by the norm-referenced approach is the relative ordering of the testees. When the adequacy of a person's performance is defined by the performances of other persons, the test is called a norm-referenced test. In norm-referenced tests, the examinee's mastery of the content covered by the test is not the point of reference.

So, depending on how it is interpreted, one's mastery of a particular skill tested could be very high while at the same time his performance relative to other persons could be very low. In contrast, one's mastery of the skill could be very low while his performance relative to other persons could be very high.

Criterion-referenced tests can be reinterpreted as norm-referenced tests (Hambleton and Novick, 1973; Hambleton et al., 1978). Such comparisons are made when the rate at which persons have mastered the skills are each compared to the absolute criterion (Thorndike and Hagen, 1977). Hambleton et al. (1978) concluded that neither the use of norm-referenced tests to make criterion-referenced measures, nor the use of criterion-referenced tests to make norm-referenced measure is satisfactory.

However, the distinction between criterion-referenced and norm-referenced measures is not easy to determine simply from the appearance of a particular instrument (test). Neither is it easy to be determined

by the particular way the scores are interpreted.

Popham and Husek (1969) suggested that the best way of recognising between the two measures is by examining

(a) the purpose for which the test was constructed,

(b) the manner in which it was constructed,

(c) the specificity of the information yielded about the domain of instructionally relevant tasks,

(d) the generalizability of the performance information to the domain, and

(e) the use to be made of the obtained test information.

Each point will become clearer as we move further through the discussion. Further distinctions have been stated by Glaser, (1963); Messick (1975); Glaser and Nitko (1971).

Hence, one should remember that discussion on norm-referenced vs criterion-referenced measurements is not about two different interpretations made about the same thing but about two different techniques of testing and test construction. Their distinction begins at the moment when the examiner decides what to test. This view was emphasised by Hambleton and Novick (1973).

4. NORM-REFERENCED MEASUREMENT

Traditionally, interpretation of most educational measures, both subjective and objective, have been based on the norm-referenced approach (Hambleton and Novick, 1973). Most standardized aptitude, achievement, personality etc. and most teacher-made tests were interpreted in a normative fashion (Martuza, 1977). Normative interpretation is to judge the adequacy of one's performance by the performances of others. This approach to interpreting test scores emphasizes discrimination among

individuals along a scale. No reference is made as to how much of the skills in a subject matter have been mastered. Scores are not high or low in an absolute sense, but are higher or lower in relation to other scores.

All measures were interpreted in terms of one of several norms. Some of these norms are the percentile norms, grade norms, age norms, and standard scores. The term norm as defined by Martuza (1977) 'refers to the statistical information which describes the distribution of scores of a well-defined population of examinees on a particular test and, provides evaluative information about an examinee's level of performance, vis-a-vis the norm population.'

Many educationists criticised the normative interpretation of scores (Thorndike and Hagen, 1977; Popham, 1978). They pointed to several problems encountered when norms are applied to test scores.

4.1. LIMITATIONS OF NORM-REFERENCED MEASUREMENT

One problem is how to specify a norming group (Angoff, 1971). The essence of the problem is that the number of possible reference groups that could be specified for each measure can become almost infinite, hence there is no way of comparing the performances of the groups. Martuza (1977) suggested that it is impractical to administer a test to very small groups or to every individual member in a reference group. He rather said that one should be satisfied if the group examined 'reasonably resemble' the parent population. However homogenous the reference group might be made, the problem of some members of the group being disadvantaged by having been grouped with a particular reference group is not going to be solved. There always remains one who is different from the rest of the reference group. Thus, any formation

of a reference group is bound to work to the disadvantage of some of its members.

Secondly, normative interpretation obscures much information about the individuals measured. It does not reveal the strengths and weaknesses of the individual in different areas of investigation (Lindvall and Nitko, 1969; Thorndike and Hagen, 1977).

Four reasons why norm-referenced tests are not suitable for educational evaluations have been summarised by Popham (1978):

(1) What is tested does not match what is taught. That is, norm-referenced tests are too general to pinpoint the effectiveness of instructional programmes.

(2) Norm-referenced tests do not provide sufficient guidance to improve instructional programmes.

(3) Norm-referenced tests are culturally biased.

(4) The fourth point which Popham (1978) calls 'Psychometric Snare' is very important. It is a trading between test information and scale construction.

Psychometricians, he argued, construct tests deliberately to spread out examinees because unless the scores of the standardization group are normally distributed there will be no way that they can compare subsequent groups. To achieve normal distribution, then, the test constructor must eliminate both difficult and easy items from the test. But how does this situation affect teaching?

One thing we know is that when a unit of instruction is taught well items sampled from it become easier and many subjects answer them correctly. We also know that items which become easy as a result of instruction are those sensitive to the instruction, and hence are good items. But according to the normative test construction criteria,

very easy items should be eliminated. Then, the best items are systematically eliminated simply because they do not conform to the normative test construction criterion. (Popham and Husek, 1969). Whenever test items sensitive to instruction are eliminated, items which are less related to the instruction remain. Then, any test composed of the latter group of items will not provide information about the effectiveness of the teaching.

Thirdly, even in selection decisions where norms are mostly applied, the same loss of information is encountered. Ebel (1962) argued that this loss of information between the score and character of performance is the main problem with the normative interpretation of scores. This is particularly true when scores from different subjects are pooled together and normalized. This interpretation of the scores does not indicate the subjects or subject areas in which the candidate did or did not do well. The best student in the school may not be the best candidate for a particular job or a programme of instruction. Selections based on normative interpretations are sometimes misleading. But this method is still the most common.

One also encounters similar problems when grade norms are used. Grade norm is the average score obtained by individuals in that grade, which has been established by administering a test to a representative sample of pupils from several grades. Then, the score of each pupil is referenced to these averages.

The problem in using grade norms is that there is no way of making meaningful inference that growth in ability between two consecutive grades. There is no way of making sure that growth in any human trait is uniform. If that it so, then it is meaningless for anyone to talk about a pupil in one grade being so many times ahead in acquisition of

knowledge of a subject than another pupil in a different grade. A 4th grader who obtained the average score of the 5th grade cannot be said to have acquired the 5th grade's knowledge. Grade norms are based on an unfounded assumption that equates intervals of time (grades) with amounts of knowledge acquired.

Secondly, the rate of acquisition of different subjects is not the same within the same individual nor is it the same between individuals. Thirdly, the use of grade norms is said to be unfair. Application of national grade norms must always favour the advantaged communities. Children from such communities usually exceed the national average. Hence, the application of norms preserves social structure where society reproduces itself (Broadfoot, 1979).

Age norms also have their problems. All human traits related to age show no uniform growth. Many such traits grow, slow down and then decline. It is an error of logic to use age norms because the use of it implies that all children are exposed to the same equal experiential opportunities, which is not true. Finally, norm-referenced measurements promote unhealthy competition and badly damage the self-concept of low-scoring individuals (Ausubel and Robinson, 1969). It is not only the low-scorers who suffer the negative effects of competition as argued by Ausubel and Robinson, but everyone. Since every student compares himself with his friend, every student's self-concept must suffer to some extent from obsession with self-aggrandisement. Even persons at the top can suffer from anxiety of keeping their present prestige. Other unhealthy effects of competition are the anxiety of the competitive situation which inhibits learning, the feeling of inadequacy, the negative climate which prevents co-operation etc.

In short, the consequence of all these is an inaccurate estimation

of the effectiveness of educational programmes. The inaccurate estimate will, in turn, lead to wrong decisions.

5. CRITERION-REFERENCED MEASUREMENTS

To overcome deficiencies encountered in norm-referenced testing emphasis in educational measurement shifted from norm-referenced to criterion-referenced measurement.

When a person's performance is compared with a criterion of proficiency, it is called criterion-referenced approach (Mehrens and Lehmann, 1975). Glaser and Nitko (1971) defined criterion referenced tests as tests deliberately constructed to yield scores directly interpretable in terms of performance standards. According to the latter definition, a test is not only interpreted differently but constructed differently in advance, to sample specific knowledge, skills and abilities. This definition gives an additional information over the first in that criterion-referenced measures differ from norm-referenced tests also in the way the criterion-referenced tests are constructed. As will be discussed later in this chapter, this means that some of the criteria used in the classical test construction are not relevant to criterion-referenced testing.

5.1. PROBLEMS OF STANDARD SETTING

Criterion-referenced tests have their own limitations too. The most difficult problem encountered in criterion-referenced testing is that of setting appropriate standards against which individual scores are to be defined. Standard is a point at which students are categorized into masters and non-masters.

Setting performance standard is the most important concept in criterion-referenced testing. The whole concept of criterion-referenced tests hinges mainly on how the performance standard is determined. Without performance standard there can be no meaningful way of interpreting scores in the CRT tradition.

Most of the literature on criterion-referenced measurement is devoted to establishing performance standards against which all individual scores are compared. There is a great diversity in the procedures used to set the standard and in the terminology used. Also, in the literature, the old controversy of subjectivity vs objectivity is renewed, in the persuit of setting an empirical standard. Some of this literature will be reviewed.

Before proceeding to the review of literature, it is worthwhile to discuss some of the difficulties faced in setting standards. The difficulties are encountered in three areas of a test. They are

(a) the appropriate difficulty of the test items,

(b) the content of the test items, and

(c) decisions related to degrees of mastery.

5.2. DIFFICULTY

However specific and well-defined the domain may be or how effective the instruction might have been, the difficulty of the task sampled can be varied as one wishes (Ebel, 1962). The sample could be very representative, but the appropriate level of difficulty of the task sampled could be anything. One may answer correctly all test items of a certain domain but cannot be said to have mastered the subject matter, when difficulty level is considered. Mastery of the subject implies that the candidate should be able to answer any question from the content of the

subject matter irrelevant of its difficulty.

In fact the problem of determining which difficulty level to include in a sample of test items can render the whole concept of mastery meaningless. One can argue that the task was too easy when many examinees reach the standard, or the task was too difficult when few examinees reach the standard. For example, a candidate who is supposed to have mastered addition tasks can still fail some addition tasks of higher difficulty level. So, one masters only a sample of tasks but not the subject matter.

5.3. CONTENT SAMPLING

When the domain is not well-defined, criterion-referenced interpretation of the scores runs into difficulty. Each candidate may do well in different areas of the same domain. Then, unless one assumes that all subdomains were of equal importance and of equal difficulty, it becomes meaningless to say that one candidate has higher mastery of the subject than another candidate. For example, when two candidates correctly answer two different sets of questions of which one of the two candidates has higher mastery, the subject becomes ambiguous.

At this point a suggestion by Hambleton and Novick (1973) is relevant. The suggestion is that a cut-off be made for each sub-scale to dichotomise examinees into two exclusive groups with regard to their mastery of each subdomain. One group is made up of those who mastered and the second group is made up of those who did not achieve the performance standard. Their suggestion means several proficiency standards for each examinee and instructional decisions to be made for each individual on the basis of his performance on each sub-scale.

Secondly, the difficulty level of the subdomains may vary. So when the difficulty level of the sets of items correctly answered by different candidates is not equivalent, it is also difficult to tell which candidate has the higher mastery. Is it the candidate who answered few but more difficult items, or is it the candidate who answered more but easier items that has the higher mastery of the subject? The criterion-referenced tradition has no meaningful solution to that problem. Of course, one can logically argue that the candidate who correctly answers more difficult questions will always answer more questions in the examination.

On the other hand, since mastery is not equated with difficulty, consideration of who answered the most difficult questions is not permissible. In criterion-referenced measurement, questions are not valued according to their level of difficulty.

5.4. DECISIONS RELATED TO DEGREE OF MASTERY

Mastery testers do not usually attach greater importance to the degrees or levels of mastery. They attach more value to the mastery-non-mastery dichotomy (Hambleton and Novick, 1973), which obscures a lot of information. For example, how meaningful is it to classify examinees at the bottom of the mastery level with examinees very close to the standard or those who just achieved the standard performance with those far above the standard?

Criterion-referenced tests acknowledge the existence of degrees of performance and its implication that decisions be made about individuals along the mastery scale. But recognition of degrees of performance poses another difficult decision to be made by the testers as to whether or not to cater for every individual along the scale or whether to group individuals arbitrarily around certain mastery levels.

6. REVIEW OF SOME PROCEDURES FOR STANDARD SETTING

Different authors used different names for the same concept. Some of these names are performance standard (Mager, 1962), criterion (Glaser, 1962), content standard test scores (Ebel, 1962), optional cutting score (Berk, 1976), etc.

All predetermined standard scores are difficult to set (Glass, 1978). Test scoring and test construction can be made easy or difficult as one wishes, so that there can be no performance standard independent of test scoring or test construction.

6.1. COUNTING BACKWARD FROM 100%

One procedure to establish criterion level is to count backward from 100% which is the desired performance level. Probably this is the most primitive procedure. The procedure recognises that perfection is not possible and the cut-off be made somewhere below perfection. But how high should the cut-off be made? Any percentage such as 80%, 95%, 90% etc. would be very arbitrary and the percentages of pupils corresponding to these criterion levels can vary greatly.

This is a pure judgemental procedure. If so, then why should one bother proposing it in the first place? Classroom teachers can state their objectives and set arbitrary mastery levels.

6.2. CONTENT STANDARD TEST SCORES

Ebel (1962) suggested two ways of securing the content meaning of the scores and avoiding the disadvantages of the normative interpretation. What he meant by securing the content meaning of test scores is setting standards against which other scores are defined.

(1) Test items are classified by judgment in their content categories. Next, each item's discriminating index is determined in the usual way. From each content category the best discriminating items are included in the test.

(2) His second method of securing test score meaning is achieved through the process of test construction. He proposed an objectively constructed test. The process of the construction is the following: Two forms of a test were constructed one by a test specialist and the other by an intelligent person with no special training in test construction, if necessary (a secretary in Ebel's experiment). Both tests were built on the basis of detailed specifications and direction. Explicit instructions were given for choosing representative sample. The two forms were administered to a sample of subjects. Half took form A and the other half took form B. The test data analysis showed that the differences in scores and item difficulty values were within the limits of sampling error (see Table 1)

TABLE 1

Analysis of Data from two forms of an objective test of word knowledge

	FORM A		FORM B
1. Student Score:			
Mean	37.3		32.8
Standard deviation	12.5		14.8
Reliability	.95		.95
Intercorrelation		.86	
2. Item Difficulty:			
Mean	11.19		8.84
Variance	79.34		72.02
Standard Error of Difference		1.24	
t		1.09	

Adapted from Ebel, 'Content Standard Test Scores'. Educational Measurement Vol. 22, No. 1, 1962.

To obtain more equivalent scores from alternate forms, tests of far more than a hundred items are required. His test was composed of one hundred words to be matched with their dictionary definitions.

Ebel's standard setting procedures may be criticised on the following points:

(1) Discriminating by difficulty is not a relevant quality in criterion-referenced test items. But he could have depended on classification of test item content by judgment as the first quality and item discrimination as a supplementary quality.

(2) The well-defined domain of word definition he used cannot be generalized to other domains.

(3) It is not clear from his experiment where one would draw the criterion cut-off.

However, Ebel's lenient approach to set criterion reflect his confidence in human judgment. Ebel (1962) said that the most objective tests rest on highly subjective foundations. The abilities, values and idiosyncrasies of the test constructor have played a major part in determining the contents of most tests. Test specification sometimes exists only in the mind of the test constructor. The process of test construction often appears to have more in common with artistic creation than scientific measurement. He said 'the quantitative sophistication of many specialists in educational measurement is displayed, not in precision and elegance of their procedures for obtaining initial measurements, but rather in the statistical transformations, elaborations and analysis they are prepared to perform on almost any raw data given them' (p. 22). The same confidence in human judgment was also expressed by Hambleton (1978). He said that the standard-setting procedure that holds most merit is that which involves judgments of test items by content specialists.

6.3. METHODS THAT USE EXTERNAL CRITERION SCORES

This is one of the procedures where the criterion score on a test is determined with the help of an external mastery (Glass, 1978). First, already judged masters are identified. The scores of the external exam taken by the masters is correlated with the criterion referenced test for which the criterion cut-off has to be established. Then, a correspondence between the masters of the external exam and the masters of the criterion-referenced tests is expected.

This technique can be criticised at least in two ways:

(a) If there is no perfect correlation between the two measures, there can be no correspondence between those who mastered the external exam and those who mastered the test. There will always remain two categories of false-negatives and false-positives.

(b) The way in which the criterion level of the external exam was first established. If one thinks the method used to determine the criterion level of the external exam was good, one should have applied the method to construct criterion-referenced tests.

Other investigators tried to establish performance standards in terms of decision-making accuracy. In that approach, criterion score is defined as the score which maximizes the probability of correct decisions and minimizes the probability of incorrect decisions (Hambleton and Novick, 1973).

6.4. Optional Cutting Scores

Berk (1976) proposed a criterion level which is selected empirically rather than subjectively on the basis of judgment. He used two equal groups of students, instructed and uninstructed. He said his method

assumes that the power of a criterion-referenced test accurately to classify students at the point where a decision is made is an indication of its quality.

Instructed students received instruction on the content and the uninstructed did not receive instruction on the content. He assumed the instructed group to possess more knowledge of the objectives than the uninstructed group. However, lack of any knowledge of any topic is difficult to assume. In most cases, some knowledge of the content must be assumed to be possessed by most of the examinees in the second group.

The test items were administered to both groups. Then criterion cut-off was made and students were divided into masters and non-masters. The criterion of classification (instructed-uninstructed) and criterion cut-off will divide the students into four possible subgroups as shown on Table 2. This is expressed graphically in Figure 1. The degree of accuracy is a function of the amount of overlap between the distributions.

The less overlap the better the test classifies. The point at which the two frequent distributions intersect is the optional cutting score. This is the score which maximizes the probability of correct decisions, P(TM) + P(TN), and minimizes the probability of incorrect decisions, P(FN) + P(FM).

7.1. Validity Coefficients

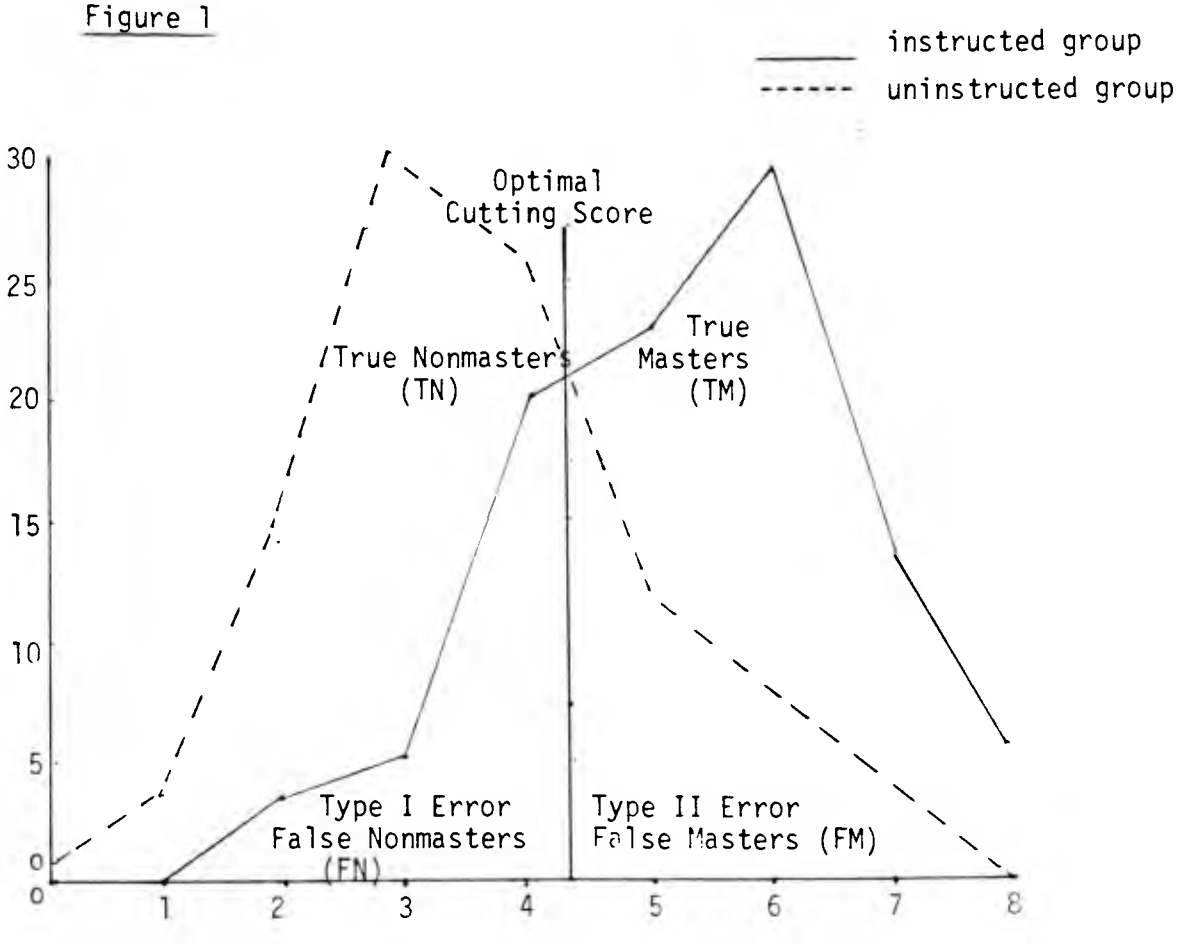
Berk's alternative procedure to derive the optional cut-off is to compute validity coefficient. Berk (1976) said 'the cut-off corresponding to the highest coefficient will yield the highest probability of correct decisions.'

TABLE 2

Predictor X Criterion Classification of Students

	CRITERION CLASSIFICATION	
	Instructed (I)	Uninstructed (U)
Predicted Masters (PM=TM+FM)	True Masters (TM)	Type II Error False Masters (FM)
Predicted Non-Masters (PN=FN+TN)	Type I Error False Non-Masters (FN)	True Non-Masters (TN)
	Masters (M=TM+FN)	Non-Masters (N=FM+TN)

Figure 1



To compute the above coefficient, he used two classifications: Predictor classification where each student at or above the prediction classification cutting score is assigned a value of 0.

Similarly, each student in the instructed group is assigned a value of 1 and each student in the uninstructed group is assigned a value of 0. A Phi coefficient, is computed between the two dichotomous variables. The cutting score is set at the score where the probability of making incorrect decisions is the lowest (see Table 3).

TABLE 3

Probabilities of Correct Decisions and Misclassification Errors, Validity Coefficient, for Cutting Scores

Cutting Score	Probability of Correct Decisions	Misclassification Errors Type II	Validity Coefficient
8	.53	.47/.00	.02
7	.58	.40/.02	.25
6	.68	.26/.06	.39
5*	.74*	.14/.12*	.48*
4	.70	.04/.26	.45
3	.58	.01/.41	.27
2	.52	.00/.48	.14
1	.50	.00/.50	.00

* Optimal cutting score

Norm-referenced tests differ from criterion-referenced tests in the context in which information provided by the tests is used in decision-making situations as well as in their relation to other aspects of the test, such as variability of scores, reliability, validity, item construction and item analysis.

7.2. VARIABILITY

Since interest is in the relative positions of individuals on a normal distribution, variability in scores is the essential characteristic in norm-referenced test interpretation. On the other hand, variability is irrelevant in the criterion-referenced tests. In criterion-referenced testing, every individual can, theoretically, obtain a perfect score, or all testees can obtain the same score etc. while the test still remains good. That kind of test would have been declared useless in norm-referenced testing.

7.3. RELIABILITY

Classical reliability varies with the degree of dispersion of scores. The wider the dispersion of the individual scores the larger the reliability coefficient would be (Nunnally, 1970; Cronlund, 1971). Hence, since the distribution of scores on criterion-referenced tests tend to be more homogenous, a low reliability index of criterion-referenced tests can be expected due to the lack of variability of test scores. So, as far as reliability is concerned, even if the test items are constructed to measure the same thing,classical reliability indices are not applicable to criterion-referenced measures. Test items should not be discarded on the basis of classical reliability procedures. According to Popham and Husek (1969), even with a negative internal consistency reliability index, a criterion-referenced test could still be good.

If classical reliability procedures, originally used to assess the reliability of criterion-referenced tests, were inappropriate to assess the reliability of criterion-referenced test scores, as Popham and Husek suggested, what alternative procedures are found to assess the reliability of criterion-referenced measures? Reliability is important in the use of any measurement and the legitimate use of criterion-referenced

measurement cannot be defended unless there is some way of assessing the psychometric properties of its measures. The advocates of criterion-referenced measurement faced the challenge and have devised new procedures to assess the criterion-referenced test reliability. The works of Carver (1970); Huynh (1976); Subkoviak (1976); Swaminathan, Hambleton and Algina (1974); Marshall and Haertel (1976), categorized as decision-consistency approaches to criterion-referenced test reliability, have been reviewed by Michael Subkoviak (1980) in 'Decision-Consistency Approaches'. The works of Brennan (1977a, 1977b, 1978); Brennan and Kane (1977a, 1977b); Kane and Brennan (in press) have been reviewed by Brennan (1980). The latter group of procedures to criterion-referenced test reliability are based on the principles of generalizability theory originally proposed by Cronbach et al. (1972).

It is not possible to give detailed technical treatment of these procedures. However, a short description of two or three procedures will be presented to give an overall picture of these new methods proposed to assess the reliability of criterion-referenced tests.

One approach to criterion-referenced reliability concerns the consistency of mastery/non-mastery decisions over repeated tests of the same subjects. The degree of consistency in classifying the mastery/non-mastery dichotomy obtained above and below a cut-off is taken as evidence of reliability of the test. That is, the number of subjects who should be the masters in the first measurement as well as in the second measurement.

Carver (1970) proposed two parallel tests to be administered to the same subjects. Then the percentages of masters on both tests to be compared. The test is reliable to the extent to which these two percentages are equal (see Table 4). Equal percentages mean perfect reliability. Though not necessarily the same individuals, 50% of the students are above

TABLE 4

Performance of Twenty Students on Parallel Ten-Item Test

	TOTAL SCORE	
STUDENTS	FORM 1	FORM 2
1	10	8
2	9	8
3	8	10
4	8	9
5	7	6
6	7	6
7	7	8
8	6	5
9	6	5
10	6	5
11	5	8
12	5	6
13	4	6
14	4	4
15	4	4
16	3	3
17	3	3
18	3	3
19	2	2
20	2	1

NOTE: Mastery Cut-Off Score = 6

---

the criterion cut-off on both forms.

However,this method can be criticised on the fact that the individuals in the two percentages may not be the same. The equal percentages procedure was insensitive to the individual and could not necessarily guarantee the reliability of the test. What is needed is a consistent procedure with which the same individual, not the same percentages,

can be classified as master/master or non-master/non-master over the repeated measures.

Hambleton and Novick (1973) proposed that the proportion of individuals consistently classified as master/master and non-master/non-master be used as index of reliability. Hambleton and Novick (1973) and Swaminathan and Algina (1974) modified Carver's method. Instead of using percentages they proposed the proportion of individuals consistently classified as master/master and non-master/non-master on two tests be used as an index of reliability (Subkoviak, 1980).

Given the same data in the above table, the second method would have rather concentrated on the number of students consistently classified. In the fictitious data, six individuals are misclassified and 14 individuals are consistently classified (see Table 5). Then reliability would be equal P = 14/20 = .70. The upper limit of the reliability index is 1.00. That is when all individuals are consistently classified.

TABLE 5

Mastery - non-mastery Outcomes on the Ficitious Data

FORM 1	FORM 2		
	MASTERY	NON-MASTERY	TOTAL
Mastery	7	3	10
Non-Mastery	3	7	10
TOTAL	10	10	20

Note: Mastery cut-off score = 6

One disadvantage of the above methods is that they require the administration of two tests. Other methods which require only one

administration were proposed by Hynh (1976); Marshall & Haertel (1976); Subkoviak (1976). The scores of the second test are simulated.

Huynh (1976) proposed a method in which the scores of the non-existing test are simulated. He computed the mean (U), variance (s^2), and KR-21 coefficient of the scores on form 1 ($\alpha 21$). Next, he computed two more parameters α and β. $\alpha = (-1 + \frac{1}{KR\text{-}21})U$, $\beta = (\quad + \frac{n}{KR\text{-}21} - n)$. These parameters plus the number of items (n) determined the shape of the joint distribution of scores on form 1 and 2, shown in Table 6 That is, when the values of α, β, and the number of scores in the test are used, in this case the scores on form 1 in Table 6. The reliability coefficient he obtained for the scores on form 1 in Table 6 and those simulated is P = .90. All the essential steps in the computation are shown on pages 134-42 in Criterion-Referenced Measurement by R.A. Berk, 1980.

Subkoviak's method. Subkoviak also proposed one test to be administered to the subjects. When his method was fitted in the scores on form 1 in Table 6 the reliability coefficient he obtained was P = .91. Steps in computing are shown on pages 143-5 in Criterion-Referenced Measurement by R.A. Berk, 1980.

Marshall-Haertel's method (1976) also requires only one administration. Using the same data in Table 6 the reliability coefficient obtained is P = .87.

8. COMPARISON OF THE APPROACHES

Subokviak (1977a, 1978) made an empirical study on the strengths and weaknesses of the above methods. He administered parallel tests of different lengths (10 items, 30 items and 50 items) to each of 1,586 students. On each test he made five criterion cut-offs at 50%, 60%, 70%

TABLE 6

JOINT DISTRIBUTION OF SCORES ON TEST FORMS 1 AND 2

	Form 2(y)										
Form 1(x)	0	1	2	3	4	5	6	7	8	9	10
0	0002	0006	0011	0013	0012	0008	0004	0002	0000	0000	0000
1	0006	0024	0050	0069	0068	0050	0028	0012	0004	0001	0000
2	0011	0050	0116	0174	0188	0152	0093	0043	0014	0003	0000
3	0013	0069	0174	0286	0338	0299	0201	0101	0036	0008	0001
4	0012	0068	0188	0338	0436	0421	0308	0169	0066	0017	0002
5	0008	0050	0152	0299	0421	0444	0354	0211	0090	0025	0003
6	0004	0028	0093	0201	0308	0354	0308	0200	0093	0028	0004
7	0002	0012	0043	0101	0169	0211	0200	0142	0077	0024	0004
8	0000	0004	0014	0036	0066	0090	0093	0072	0040	0014	0003
9	0000	0001	0003	0008	0017	0025	0028	0024	0014	0006	0001
10	0000	0000	0000	0001	0002	0003	0004	0004	0003	0001	0000

NOTE: Each entry in the body of this table represents the proportion of examinees that would obtain score x on form 1 and y on form 2. Decimal points are omitted.

This table was taken from: Criterion Referenced Measurement by R.A. Berk, 1980.

and 80% of the items correct. (See Berk, 1980).

The relative advantages and disadvantages are summarised in Table 7.

TABLE 7

Relative Advantages and Disadvantages of the Reliability

METHODS	ADVANTAGES	DISADVANTAGES
Swaminathan	Computationally simple, unbiased estimates.	Requires two testings; large error of estimate for small groups.
Subkoviak	Requires one testing; small errors of estimate.	Computationally tedious, biased estimates for short tests.
Huynh	Require one testing; small errors of estimate.	Computationally tedious (except for approx.); biased (but conservative) estimates for short tests.
Marshall-Haertel	Require one testing; small errors of estimate.	Computationally tedious; biased estimates for short tests.

This table was adopted from R.A. Berk (1980) Criterion-Referenced Measurement. Baltimore: John Hopkins Univ. Press, 1980.

9. VALIDITY

Due to lack of variability and hence reliability on the criterion-referenced tests, the usual validation procedures which make use of correlational techniques to determine validity coefficients are not applicable (Hambleton, 1980) to determine the validities of criterion-referenced tests. Methods of content or construct validities can be more appropriate in criterion-referenced tests (Popham and Husek, 1969; Hambleton, 1979). Since criterion-referenced test is a measure of achievement, content validity is more important for criterion referenced

measurement (Mehrens and Lehmann, 1975). Approaches to criterion-referenced validation are discussed under content validity and elsewhere in the study.

10. ITEM CONSTRUCTION

As long as the items represent the behaviours delimited by the criterion, too easy or too difficult test items are permissible in criterion-referenced testing. On the other hand, the norm-referenced test item writer would have avoided such items, which do not discriminate betwen testees and contribute nothing to the test variance.

TABLE 8

Purposes for Criterion-Referenced Tests

FOCUS OF THE TESTING PROGRAMME	PLANNING	TYPE OF DECISION	
Student	Diagnosis, prediction and placement	Determination of mastery grades and success of placement	Instruction between the student, the group and the programme
Group, classroom, ethnic, ses,cult. geog. groups etc.	Classroom management, curriculum selection	Instructional and administrative accountability	Interaction between groups and programmes
Programme	Organisation and sequencing of instruction curriculum and product develop. needs assessment	Programme evaluation analysis of subject matter domain	Comparisons between types of programmes, analysis of programme components develop. of measurement methodology.

11. CONCLUSIONS

Norm-referenced measurement experts offer no solution to the problem of specifying an appropriate norming group, the loss of information about examinees' absolute performance, the unfair discrimination, the fierce competition etc.

Nobody should deny the limitations of norm-referenced measurement. But what one would demand, if he is to reject NRT, is an empirical evidence of the extent to which these limitations affect the usefulness of a test and whether or not better alternatives are available.

In criterion-referenced measurement, the problem encountered in setting standard corresponds to that of specifying an appropriate norm. No satisfactory empirical procedure has yet been established. Most of the procedures so far proposed depend on human judgment (Angoff, 1971; Ebel, 1979; Jaeger, 1978), or have complex computational requirements.

With regard to social consequences of norm-referenced testing, such as competition, cultural bias etc., one should not be deceived to believe that criterion-referenced testing offers a satisfactory solution. For example, in the case of competition, if the student is not directly competing with others he must be competing with time. Trying to master a task earlier than others is itself a form of competition. Those who master the task earlier must have more opportunities than those who master it later, as those at the top of a normative scale have more opportunities than those who are at the bottom of the scale.

Norm-referenced testing is criticised to favour examinees from advantaged communities but it cannot also be ruled out in CRT either. Candidates from advantaged backgrounds will always do better than others and hence will be accorded a better treatment. Classifying individuals who are far apart, but on the same side of the criterion cut-off, into the same mastery category is unfair and misleading.

CHAPTER 5

TEACHERS AND PSYCHOMETRICIANS' OBJECTIVES IN ADMINISTERING TESTS AND EXAMINATIONS

1. INTRODUCTION

Given tne problem of determining how much of the objective psychometric procedures can be safely left to teachers, it was necessary to reformulate the issue in more testable terms. The first question for research concerned the objectives of assessment as seen by professional teachers and psychometricians. Are the objectives the same for both groups? Without a clearly defined notion of the objectives it is not possible to evaluate the achievement, nor is it possible to compare the methods.

The first stage of the research is devoted to identifying what teachers expect from tests and assessments. It was therefore decided to review the known literature of teachers' attitudes and opinions of psychometric tests and to follow this with some empirical investigation of one's own.

This chapter reviews the work done on teachers' attitudes and opinions on psychometric tests, then proceeds to elicit opinions from teachers about the relative merits of both methods of assessment. The opinions are obtained on attitude type scales. Each item in the scale invites the teacher to express a preference for one or the other method.

A judgment is therefore necessary. The research interest is in determining the reasons for these judgments. By appropriate analysis of these judgments, it should be possible to isolate the number of dimensions professional teachers use in evaluating tests in general. These dimensions will provide an insight into teachers' criteria and expectations for testing and assessing pupils.

2. THE BACKGROUND LITERATURE TO TEACHER OPINIONS AND ATTITUDES TOWARD PSYCHOMETRIC TESTS

The existing literature on teachers' attitudes toward psychometric procedures of testing and test constructions compared to professional procedures of testing and test constructions is very sparse. Goslin (1967) asked public secondary school teachers, private secondary school teachers, secondary school counsellors, elementary school teachers and elementary school principals, how accurate they (personally) felt most standardized intelligence or aptitude tests were in measuring a student's potential. Of course, he asked the question in comparison with other subjective measures such as teachers' evaluations, non-standardized tests, parents' opinions etc.

The results showed that over 70% of all the respondents felt that most standardized tests are much more accurate than other measures. Only 20% of the respondents in the study felt that standardized tests are not more accurate than other measures. About 5% of the respondents in the study felt standardized tests are not as accurate as other measures. Less than 1% felt that standardized tests are much less accurate than other measures. This pattern of response reveals that a great majority of all the respondents felt that standardized tests are better indicators of students' intellectual ability and academic achievement than other measures.

A second question asked was which of several measures (7 of them) commonly used provided the most accurate single measure of the students' intellectual ability? These measures were the grade point average, parent opinion, standardized achievement test scores, intelligence or scholastic aptitude test scores, teacher opinion, student's own opinion of his ability and peer opinion. Intelligence or scholastic aptitude test scores

were indicated by respondents to be more accurate than all other measures with around 38-57% of the respondents, across categories of respondents. Teachers' opinion and standardized achievement test scores received nearly an equal percentage of endorsements from the respondents,and came after intelligence or scholastic aptitude test scores with endorsements varying from 10-30% across respondent categories. Third came the grade point average varying from 8-17% of the respondent categories. At the bottom of the list came the parental opinion of students' intellectual ability, receiving less than 1% of the responses in any category.

As far as accuracy is concerned, Goslin concluded that teachers believed standardized tests to be the most accurate measure of a student's intellectual potential and academic achievements.

A third question he asked was whether they thought that teachers should consider their pupils' intelligence test scores in assigning grades. 'Do you think that teachers should consider their pupils' standardized achievement test scores in assigning grades in their courses?' Of those who expressed their opinions 3% said that they always considered intelligence test scores when assigning grades to pupils; 23% considered them frequently; 40% considered them in special cases; 40% never considered them at all.

In replying to the item dealing with standardized achievement test scores as distinct from I.Q. tests, the teachers revealed that 7% always considered ahievement test scores when assigning grades in their classes; 13% frequently considered them in special cases only; 41% never considered them at all.

Of the two standardized tests, intelligence or scholastic aptitude,

tests were regarded the most accurate single measure of the student's intellectual ability and the most important factor to consider when assigning course grades to pupils.

Whether teachers perceived intelligence test scores as more objective than achievement test scores, and therefore believed them to be more accurate, was not clear from the teachers' responses.

Different groups of respondents expressed varying degrees of confidence in different measures. That fact has been reflected in their attitudes. For example, secondary school counsellors expressed the greatest degree of confidence in objective measures.

The teachers who expressed greater confidence in objective tests were those who had greater familiarity with tests. By familiarity with tests, or psychometric sophistication, the author means taking one or more major standardized ability and/or achievement test once or more in one's life-time; administering these tests to subjects; reading copies and examining their contents; hearing about them;/having taken one or more psychometric courses.

In contrast to these groups are secondary school counsellors, and elementary and private school teachers who expressed relatively less faith in the accuracy of more objective measures (see Goslin, 1967, tables 25-26). About 55% and 64% respectively of the latter two categories of respondents reported to have taken no course in tests and measurements. They also reported limited familiarity with tests in other ways.

As psychometric tests are associated with school psychologists, teachers' attitudes towards school psychologists may reflect their attitudes towards psychometric tests. In both Goslin (1967) and in Kessler et al. (1973), teachers expressed high positive attitudes towards

psychometric tests (Goslin) and towards school psychologists (Kessler et al).

However, whether or not these positive attitudes expressed by the teachers towards the school psychologists can be taken to indicate their positive attitudes towards psychometric tests was not discussed by Kessel and his colleagues.

Among other questions, Romig (1970) investigated teachers' attitudes towards school psychologists. He asked :-

(a) whether the attitudes of classroom teachers towards school psychologists were generally positive or negative;

(b) whether there were correlates of the teachers' attitudes.

Even though his subjects expressed dissatisfaction with the services the psychologists provided, the majority of the teachers believed that the services of the school psychologists were needed.

Male teachers responded more positively to testing and school psychologists than female teachers. Humanities teachers were more positive than science and business teachers. Lower grade level teachers were more positive than the higher grade level teachers.

From the scant research in this area, it seems that a large proportion of teachers regarded standardized psychometric tests as positive aids in assessing pupils. However, a sizeable proportion hold the contrary view (41%). Even more revealing is the fact that respect for these tests varied with the level of teaching and the type of schools.

It appears that the efficiency and the usefulness of psychometrically constructed tests are still questioned by a large number of professionals. There is therefore good grounds for such an investigation into the reason for such a high proportion of disagreement. One would also like to know whether these teachers hold the same views regarding the purpose

of the assessment as those psychometricians who construct and administer these tests.

3. METHODOLOGY

As indicated previously, the basic design required a method of eliciting a judgment from teachers. An attitude type scale was chosen to be the most appropriate method. A factor analytic technique was decided upon as a suitable way of uncovering the basic dimensions used by the teachers in making their judgments. Consideration was given to cognate techniques like repertory grids, but the final decision was based on a concern for objectivity.

3.1. SUBJECTS

The scale described below was administered to 65 male and female U.K. teachers. They varied in duration of teaching experience and had taught different subjects at different levels. Some had taken one or more coureses in educational and psychological measurement. The teachers also differed in respect of the size of the classes and the age range of the pupils varied from 8 to 35 years.

3.2. MEASURING INSTRUMENT

The scale described records teachers' opinion towards the relative efficiency of two procedures for testing pupils. The two procedures compared are:

(a) the psychometric method of testing and

(b) the professional method of testing pupils.

The reason for including reference to two objects of methods in the same scale is to elicit a judgment of preference to each item. In

each case the respondent uses a criterion to decide which of two things is better. No response is given by a subject without the person having something else in mind relative to the present stimulus (Thurstone, 1929). Since the research interest of this study is in the reasons for the decision rather than the decisions themselves,a sufficient number of decisions had to be elicited.

The first section of the questionnaire, as shown in Appendix I, deals with background information about the subjects. The subjects were requested to state their sex, years of teaching experience, subject taught, the number of years teaching the subject, the number of years teaching the present class of pupils, the approximate size of the class, its age range and whether the teacher had taken a course(s) of educational measurement. This section also includes a statement of our aims and two definitions of the two procedures mentioned above.

The second part of the questionnaire consisted of instructions to the subjects, two examples, 54 attitude statements and their response categories (See full text in Appendix I).

The statements were arranged in a Likert-type scale with five response categories that could be answered: strongly agree; agree; uncertain; disagree and strongly disagree. The likert method of attitude scale has been chosen because the technique produces a more homogeneous scale, allows a degree of intensity of sentiments and requires no judges to sort statements (Mehrens and Lehmann, 1975).

There are an approximately equal number of affirmative and negative statements. Every attempt was made to assure the objectivity and validity of the scales.

3.2.1. BREAKDOWN OF THE UNIVERSE DOMAIN

The universe of content from which the attitude statements have been sampled was broken down into 6 subdomains:

(1) The first domain deals with accuracy. The set of statements referring to the accuracy dimension of the domain were further subdivided into:

(a) those statements (8 statements) referring to the accuracy of psychometric tests;

(b) those statements (8 statements) referring to the accuracy of professional tests;

(c) those statements (9 statements) referring to the accuracy in construction of either the psychometric or professional tests and

(d) those statements (4 statements) referring to the accuracy of prediction of either type of test.

Hence there are 29 items in the accuracy domain.

(2) The second set of statements (9 statements) refer to fairness in tests. By fairness, one refers to the ability of the method of assessment used to discriminate subjects only according to the degree to which each possesses the individual trait measured. This point has been discussed extensively by Jensen (1980).

However, fairness of a test, as perceived by teachers, pupils and parents does not necessarily mean the same as defined by psychometricians. To many people test fairness may have limited sense. Their judgments of a test's fairness being based only on their belief about particular individuals being evaluated.

(3) The third set of statements (5 statements) refer to the adequacy of either procedure in sampling the content.

(4) The fourth set of statements (4 statements) refer to cost or convenience of either assessment procedure in testing and/or test construction.

(5) The fifth set of statements (4 statements) refer to whether or not teachers believe that psychometric tests and testing are not part ofthe normal function of the teacher, but belong to professional psychologists.

(6) The sixth set of statements refer to the test's ability to facilitate interpretations of test results among the professionals as well as among others.

As shown above, the subdomains were not equally represented in the scale. We believed that some were more important than others.

The statements were arranged in such a way that statements which belonged to one subdomain or area were scattered throughout the scale (see Mehrens and Lehmann, 1975).

Each statement from one area was placed one or more steps away from the next statement belonging to the same area. That was to avoid interdependence of item responses. The pattern of this arrangement of the 54 statements is shown below:

1(a);3;1(d);1(c);2;1(b);2;1(a);3;1(c);2;1(b);2;1(a);4;1(c);5;1(b);
2;1(a);3;1(d);1(c);5;1(b);2;1(a);3;1(c);5;1(b);2;1(a);3;1(d);1(c);
5;1(b);2;1(a);4;1(c);7;1(b);2;1(a);4;1(c);6;1(b);2;1(c);4;1(d).

Each of the above 54 codes stands for a statement. The codes indicate the subdomains and the positions of the statements they represent. Each code was in turn assigned a number. So, the first 1(a) ofthe arrangement for example, corresponds to the first statement of set (a) statements of subdomain 1, its accuracy, and at the same time

1(a) corresponds to the first statement of the list of attitude statements.

Statements numbered 1,8,14,20,27,33,40 and 46 belong to set (a) statements of subdomain 1, accuracy; statements numbered 6,12,18,25, 31,38,44 and 50 belong to set (b) statements of subdomain 1, accuracy; statements numbered 4,10,16,23,29,36,42,48 and 51 belong to set (c) statements of subdomain 1, accuracy; statements numbered 3,22,35 and 54 belong to set (d) statements of subdomain 1, accuracy; statements numbered 5,7,11,19,26,32,39,45 and 52 belong to subdomain 2, fairness; statements numbered 2,9,21,28 and 34 belong to subdomain 3, content sampling; statements numbered 15,41,47 and 53 belong to subdomain 4, convenience or cost; statements numbered 17,24,30 and 37 belong to subdomain 5, separate provision for tests and testing; statements numbered 43 and 49 belong to subdomain 6, ease of interpretation. The actual statements are shown in Appendix I.

3.2.2. OTHER STEPS TAKEN TO GUARD AGAINST BIAS

However confident one might be that his respondents did not give biased responses, there is no absolute guarantee that people's attitudes and their actions do not contradict each other. One may give a verbal expression of one thing and do something different. Here we have taken some more steps to eliminate factors which could cause biased responses.

Firstly, we checked our statements against suggested criteria. The criteria against which we have checked our statements as recommended by such authorities as Allen Edwards (1957); Thurstone and Chave (1929); Likert (1932) and Oppenheim (1966), are as follows:

1. Avoid statements that refer to the past rather than to the present.
2. Avoid statements that are factual or capable of being interpreted as factual.
3. Avoid statements that may be interpreted in more than one way.
4. Avoid statements that are irrelevant to the psychological object under consideration.
5. Avoid statements that are likely to be endorsed almost by everyone or by almost no-one.
6. Select statements that are believed to cover the entire range of the affective scale of interest.
7. Keep the language of the statements simple, clear and direct.
8. Statements should be short.
9. Statements containing universals such as all, always, none and never,often introduce ambiguity and should be avoided.
10. Each statement should contain only one complete thought.
11. Words such as only, just, merely and others of similar nature should be used with care and moderation in writing statements.
12. Wherever possible, statements should be in the form of simple sentences rather than in the form of compound or complex sentences.
13. Avoid the use of words that may not be understood by those who are to be given the completed scale.
14. Avoid the use of double negatives.

3.2.3.PRE-PILOTING QUESTIONNAIRE

Before administering the questionnaire to subjects the reactions of several representative teachers (10) were sought. They were asked to read the whole questionnaire and indicate any difficulties which they

thought might arise due to ambiguity which would not be understood by their fellow teachers when they attempted to fill the questionnaire; due to the task being boring for the respondents; or due to the language used not being simple enough and straightforward, so that ideas conveyed in the statements could not be easily comprehended.

3.2.4. Results of the Pre-Piloting

Among other things, the teachers suggested that the meaning of statistical or psychometric procedure or tests should be defined clearly in advance. Instead of statistical or psychometric procedure or psychometric tests, they believed that test psychologists, school psychologists,psychological tests or standardized tests are terminologies more widely understood by most teachers. They also reported, with less unanimity, some minor difficulties that could be encountered by the subjects. These minor difficulties included:

(a) that the respondent's name should not be asked in the questionnaire;

(b) that the present questionnaire was a bit long for volunteer subjects to fill; and

(c) the language used should be made simpler.

3.2.5. Modifications

The questionnaire was accordingly modified in the light of these suggestions. According to their suggestions the meanings of both statistical or psychometric procedure of test construction, and professional or subjective procedure of test construction have been defined as shown on the next step. Some of the statements were rewritten to satisfy some of these suggestions. The number of items containing affirmative and negative statements were balanced.

The scale assured complete anonymity of the identity of the respondent. Apart from these steps taken to guard against bias, the nature of the attitude object tests seemed not to be evocative. On the whole, one could place reasonable confidence in the statements that they evocated no biased reaction of the respondents that could affect the results.

3.3. THE PROCEDURE

The questionnaire was presented to the subjects both individually and in groups, depending on their availability. Next, the subjects were instructed to read the definitions of psychometric and professional procedures of test construction. What we mean by psychometric procedure of test construction is that process in which test items have been written, pre-tried on a sample of subjects and then subjected to a statistical analysis, so that all test item properties such as difficulty, discrimination, validity, reliability etc. for each item, become known. Secondly, by professional procedure of test construction, we meant a classroom teacher-made test or teacher's own personal ratings of the pupils, which have not been subjected to statistical analysis, so that no test item property is known in advance. Subjects were told that the aim ofthe study was only to know when and under what conditions teachers would prefer to use psychometric procedures to professional procedures of assessing pupils in a class, or vice-versa.

Next, each subject was asked to report on the age of the children in his/her class, educational level, linguistic ability, social background etc. In each statement, the subject was asked to endorse a response category of his choice on a five point scale by ticking a blank space provided. The responses required were whether he strongly agreed, agreed, was uncertain, disagreed or strongly disagreed with the statement. Subjects were given enough time to finish the task.

CHAPTER 6

ANALYSES AND THE RESULTS OF THE ATTITUDE QUESTIONNAIRE

1. PURPOSES

The purpose of this questionnaire was to construct a measure of teachers' judgments about psychometric vs professional testing; to select a smaller number of items which have the highest loadings on the attitude and to eliminate all items that do not belong to the attitude. Another aim of the attitude questionnaire was to give evidence on teachers' attitudes toward psychometric testing, and to indicate which background variables were related to the teachers' positive attitudes (if any) toward the psychometric testing.

2. ITEMS ANALYSIS

Items were coded so that a high score reflected a positive attitude toward psychometric testing. On a five point Likert-scale, statements favouring psychometrics received 5 points for strongly agreeing, 4 points for agreeing, 3 points for uncertainty, 2 points for disagreeing, and 1 point for strongly disagreeing. Statements which did not favour psychometric testing received 1 point for strongly agreeing, 2 points for agreeing, 3 points for uncertainty, 4 points for disagreeing and 5 points for strongly disagreeing.

3. CRITERIA OF ITEM SELECTION

Analysis of the attitude scale was carried out, Item-total statistics were computed for the scale. With one item deleted, the mean, the variance, the item-total correlations of the scale, squared multiple correlations and alpha were printed out.

The most commonly used criteria for item selection is the item's correlation with the total score on the test and the item's factor loading on the desired factor. Item-total correlation is the relationship between individual items and the rest of the items as a set. 21 items were selected from the attitude scale on the basis of these two criteria. The items selected for each factor were those with the better item-total correlations with the scale (Mehrens and Lehmann, 1975). Secondly, the items which had the highest factor loadings with their respective factors. Items from the selected factors which did not correlate significantly with the total score or did not have significant factor loadings with their respective factors were discarded (Jensen, 1980; Guilford and Fruchter, 1978).

With a sample size of 60 and above correlation coefficients of .25 and above are significant at the 5% level and correlation coefficients of .33 and above are significant at the 1% level (Child, 1970). The factor loadings of all the 21 items included in the final scale are significant at the 1% level.

4. THE REVISED SCALE

A new scale was created from the 21 items by adding the scores of 65 subjects on these 21 items. About 20 items of sufficiently high alpha (.80) are usually recommended for final attitude scales (Nunnally, 1978; Oppenheim, 1966). Then,

(1) t-test was carried out to test whether or not the means of the male and female groups in the scale significantly differed from each other.

(2) Another t-test was carried out to test whether or not the mean of those subjects who have taken one or more courses in educational

and psychological measurements significantly differed from the mean of those who had none.

(3) One-way analysis of variance was carried to test whether teachers' attitudes differed according to the subjects they have taught.

The analysis has shown that items 8,10,21,27,34,35,38 and 40; 6,18,23,25,31,33 and 47; 5,16,17,29,35,44 and 49 have better significant correlations with the scale than other items of the same respective factors. The same items have the highest factor loadings on their respective factors. As mentioned earlier, the criteria of item selection for the new scale was based on two results, the factor loadings and the item-total correlations. So, items with the best item-total correlations and the highest factor loadings are included in the revised scale.

5. FACTOR ANALYSIS

One aim of using factor analysis was to abstract fewer factors from the larger number of variables (Cattell, 1978; Kim and Mueller, 1978; Oppenheim, 1966). Factor analysis is also used to determine the internal statistical structure of a set of items (variables) proposed to measure a construct (Nunnally, 1978) and to examine the dimensionality of attitude scales (Oppenheim, 1966). Both were aims of the attitude questionnaire.

The scores made by the subjects on the scale were factor analysed to abstract the underlying dimensions. The purpose of factor analysis is either to confirm a hypothesis about the number and the kinds of factors underlying the data, or to explore the number and the nature of factors underlying the data without holding any particular hypothesis (Nunnally, 1978). But whether or not one has a previous hypothesis about the number and the nature of the factors is irrelevant to the statistical

6. REVISED SCALES

TEACHERS' ATTITUDES TOWARDS PSYCHOMETRIC VERSUS PROFESSIONAL TESTING PROCEDURES

I am conducting research into teachers' feelings about psychometric as opposed to teachers' own tests. Would you be kind enough to assist by completing the following questionnaire.

Information about Teacher

1. Sex: Male ____ Female ____
2. Number of years in the profession: _____years
3. (a) Subject taught: Languages ___, Social Science ___, Maths ___, Science ___, Others ___, _____, _____, _____.
   (b) How long? ___, ___, _____, _____, _____. ____, ____, _____, _____, _____,
4. Age range of the pupils taught at present: ___ to ___
5. How familiar are you with the following:
   (a) Constructing psychometric tests: Very familiar _____ Fairly familiar _____ Not at all _____
   (b) Administering psychometric tests: Very familiar _____ Fairly familiar _____ Not at all _____

Our aim is to know when and in what condition(s) teachers incline more toward professional or psychometric procedures of assessing pupils in a class. What we mean by psychometric procedure of test construction is the process in which test items are written, tried out on a representative sample of subjects and then submitted to statistical analysis. By professional, we mean teacher-made tests or teachers' own personal ratings of the pupils which have not been subjected to statistical analysis.

Now, suppose you were asked to measure a given ability or characteristic of children in one of your classes, say their arithmetic attainment. Assume that you have knowledge of the children's age, educational level, linguistic ability, social background etc. Keeping that class of children in mind, would you, please, answer the following statements by ticking the appropriate column: Strongly agree, agree, uncertain, disagree, strongly disagree. For example, someone in favour of the first statement below but feels that there are some exceptions would tick as follows:

Examples	Strongly Agree	Agree	Un-certain	Disagree	Strongly Disagree
(a) Children bring a husband and wife closer to each other					
(b) On balance, children are more of a blessing than a burden					

Similarly, someone who does not particularly like children could disagree with the second statement and tick the strongly disagree column instead.

The Attitude Scale

STATEMENTS	RESPONSE CATEGORIES				
	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
1. Given two scores of a student's academic ability, the most accurate assessment of the student would be to assign more weight to the assessment of the psychologist than the professional assessment of the classroom teacher.					
2. I think the single most accurate measure(test) of a student's intellectual and academic ability is the psychologically constructed test of the psychologist.					
3. A great deficiency in teacher-made tests, compared with standardized tests, is that there is no way of determining, in advance, whether the test (teacher-test) is too difficult or too easy for the students.					
4. On the whole, psychological tests are not more accurate than personal judgments.					
5. Inaccuracy is the main fault of classroom teacher-made tests when compared with the tests constructed by psychologists.					
6. Psychological tests are less accurate than classroom tests or teacher judgments.					
7. I feel that an experience of two or more years with the class results in a more accurate estimate of students' academic attainments than a psychological test.					

STATEMENTS	RESPONSE CATEGORIES				
	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
8. I believe that I know my students better than any test can tell.					
9. Since their test items will be anyway subjected to rigorous item analysis later on, psychologists should not worry about the initial construction of the test items.					
10. Without using any test, I could rank nearly all my class according to their knowledge of the subject I teach					
11. I believe I could assign overall grades to my pupils without giving them a written test.					
12. I believe there is no harm in assigning course grades partially on the basis of earlier standardized test scores of the students.					
13. Unless compelled otherwise I will use essay type questions most of the time.					
14. I trust my own personal ratings of my class less than I would trust a test constructed by a psychologist.					
15. It is unfair to use personal judgments as a measure of students attainments since they are influenced by other characteristics of the pupils.					

STATEMENTS	RESPONSE CATEGORIES				
	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
16. I believe all tests items should be subjected to statistical analysis before I administer them to a class.					
17. Each school should have its psychologist responsible for only testing and test construction					
18. No valid test can be constructed without following a table of test specification					
19. We must always substantiate subjective judgments in teacher-made tests by statistical analysis of the test items.					
20. In an essay type test since candidates do not attempt exactly the same questions, it is less objective to compare the performances of the pupils in the same class.					
21. Teachers' evaluation tools such as teacher-made tests are as objective as psychological tests when properly constructed.					

procedure. In this questionnaire, there was no definite hypothesis made in advance about the number and the nature of factors to be extracted from the data. The aim of the factor analysis in this questionnaire was mainly exploratory.

7. THE RESULTS OF THE FACTOR ANALYSIS

In the first phase of the factor analysis, the number of factors printed was equal to the number of variables. However, the results of the factor analysis have shown that only a few factors accounted for most of the variance (see Table 1). One factor alone accounted for more than 20% of the variance. Half of the 54 factors printed accounted for less than 8% of the variance. Some accounted for zero of the variance.

The factor analysis was repeated. This time only 6 factors were requested to be printed. The results of the second analysis are as shown in Tables 2+3

When the 6 factors were requested, the above percentages were printed. This time, factor 1 has drawn relatively more of the variance left by the smaller factors which have been eliminated by the procedure. The 6 factors were examined whether or not they represented the 6 subdomains into which the attitude scale has been broken down (see Chapter 5, page 73). However, the factors printed were thought not to have reflected the original domains. The numbers of all the items and their factor loadings on the 6 factors are shown in Table 2

There are two common criteria for determining the number of factors to be extracted from a factor analysis. According to Kaiser's criterion, only factors with latent roots greater than one are considered as common factors (Child, 1970). If employed, this criterion would give us about

TABLE 1

Factors, Percentage accounted for by each Factor and Cumulative Percentages

FACTOR	PERCENTAGE	CUMULATIVE PERCENTAGES
1	20.3	20.3
2	8.1	28.4
3	5.7	34.1
4	5.4	39.5
5	4.9	44.4
6	4.1	48.5
7	3.9	52.3
8	3.5	55.9
9	3.3	59.2
10	3.3	62.5
11	2.9	65.3
12	2.7	68.0
13	2.4	70.4
14	2.4	72.8
15	2.3	75.1
16	2.1	77.2
17	2.0	79.1
18	1.7	80.9
19	1.7	82.5
20	1.6	84.1
21	1.4	85.5
22	1.4	86.9
23	1.3	88.2
24	1.2	89.4
25	1.1	90.5
26	1.0	91.5
27	1.0	92.5
28	.9	93.4
29	.7	94.1
30	.7	94.8
31	.6	95.4
32	.6	96.0
33	.5	96.5
34	.4	97.0
35	.4	97.4
36	.4	97.8
37	.4	98.2
38	.3	98.5
39	.3	98.8
40	.2	99.0
41	.2	99.2
42	.1	99.3
43	.1	99.4
44	.1	99.5
45	.1	99.6
46	.1	99.7
47	.1	99.8
48	.1	99.9
49	.0	99.9
50	.0	100.0
51	.0	100.0
52	.0	100.0
53	.0	100.0
54	.0	100.0

TABLE 2

Factor Loadings on the First Six Factors

VARIABLES	FACTORS					
	1	2	3	4	5	6
1	.50				.50	
2	.60					
3						
4		.59				
5		.60	.50			
6	.44		.56			
7	.46					
8	.88					
9		.50				
10	.64			.40		
11	.40					
12	.49				.41	
13				.40		
15	.50		.41			
16		.49	.40			
17		.63	.40			.41
18						
19	.56					
20	.64					
21	.68					
22	.68					
23			.41			
24				.40		
25		.53				
26						
27	.61					
28			.44			
29						
30						
31		.42				
32	.41					
33						
34	.61					
35	.70					
36						
37						.43
38	.73					
39						
40	.78					
41	.54					
42				.40		
43	.48					
44		.47				
45	.43					
46	.52					
47		.40				
48						
49						
50	.58					
51	.59					
52						
53					.41	
54	.59					

TABLE 3

Percentages of the variance accounted for by each of the first six factors and the cumulative percentages

FACTOR	PERCENTAGE	CUMULATIVE PERCENTAGES
1	46.6	46.6
2	16.1	62.7
3	11.2	73.9
4	9.6	83.5
5	9.4	92.9
6	7.1	100.0

TABLE 4

Percentages of the variance accounted for by each of the first three factors and the cumulative percentages

FACTOR	PERCENTAGE	CUMULATIVE PERCENTAGES
1	64.6	64.6
2	20.7	85.3
3	14.7	100.0

TABLE 5

Factor Loadings on the first three factors

Variables	FACTORS		
	1	2	3
1	.42		
2	.60		
3			
4			
5			.49
6		.56	
7	.40		
8	.86		
9			
10	.62		
11			
12	.46		
13			
14			
15	.60		
16			.49
17			.68
18		.30	
19	.60		
20	.62		
21	.65		
22	.60		
23		.40	
24			
25		.60	
26			
27	.60		
28			
29			.44
30			
31		.53	
32	.31	.43	
33		.41	
34	.70		
35	.70		.40
36			
37			
38	.60		
39			
40	.80		
41	.41		
42			
43	.44		
44			.47
45	.47		
46	.41	.54	
47		.40	
48			
49			.40
50	.53		
51	.62		
52			
53			
54	.53		

17 factors. The second criterion is Cattell's Scree test. In this technique, the latent roots are plotted against the number of factors (see Figure 1). The point at which the curve straightens out is taken as the maximum number of factors to be extracted (Child, 1970). This criterion has been used here.

Looking at the Eigenvalues and the percentages of variance printed for each factor, the main departure seems to have occurred between factors 3 and 4. Then, a 3-factor solution has been requested and the percentages in Table 4 were obtained for the three factors.

These 3 factors have been interpreted and named as :

(1) accuracy,

(2) prior knowledge, on the part of the assessor, of the persons to be assessed and the subject matter to used for the assessment, and

(3) the objective preparation, the extent to which the test has been carefully and objectively planned before it has been answered.

The three most important dimensions thought to have underlain the responses of the subjects have been interpreted as above. The accuracy dimension is the main dimension underlying the responses of the subjects. More items have higher loadings on that factor.

The second dimension concerns whether or not a proposed technique of assessment is being designed, in advance, with regard to a prior knowledge of the subjects to be assessed and the subject matter to be used for the assessment. It has been interpreted that a prior knowledge of assessees and the subject matter has been regarded by the respondents as a desirable quality to be incorporated into the test construction process. That a method of assessment which does not consider the age, educational

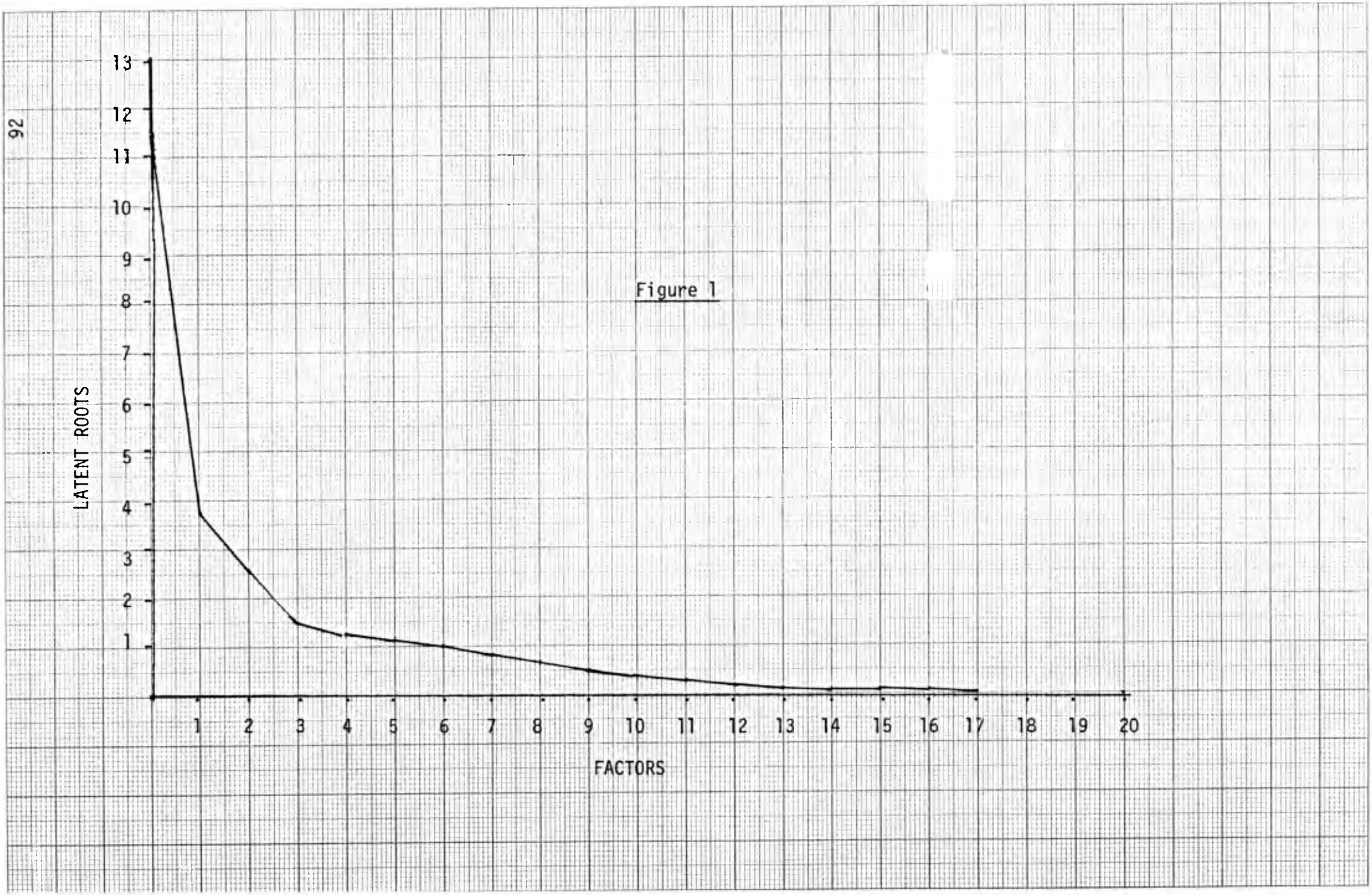

LATENT ROOTS
13
12
11
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
FACTORS

Figure 1

level, cultural background, etc., of the persons to be assessed can never be appropriate for measurement. Therefore, before constructing a technique of assessment, one must keep in mind the characteristics of the population to be assessed and the material to be used for the assessment.

In the third dimension, objectivity refers to a well thought and better organized manner of approaching testing and test construction activities. Objectivity is not equated with empirical evidence. It does not matter whether one employs an empirical technique or human judgment. With either method, one can be more objective or less objective. To be objective is to consider all relevant factors in the preparation of testing and test construction. To be less objective, on the other hand, is to overlook many relevant factors in testing and test construction situations.

This factor relates to that which preceded it. One can take relevant factors in testing and test construction into consideration only when one has prior knowledge of the assessee, the material to be used for the assessment, and the conditions of assessment. When these dimensions are combined they produce a third dimension, accuracy. A technique of assessment is accurate to the extent to which these two dimensions are maximized.

From each of the three factors, the items (7 from each factor) with the higest factor loadings have been selected to be included in the final scale (see p.82 and Appendix I). Items which have significant factor loadings on more than one factor were avoided.

8. THE RESULTS OF THE T-TESTS AND THE ANALYSIS OF VARIANCE

8.1. THE RESULTS OF THE T-TESTS

A **t**-test was carried out to compare the means of the male and female subjects in the questionnaire. There were 35 male subjects and 27 female subjects. The means of the two groups were 69.89 and 65.30 respectively. The difference between the two means was not significant (see Table 6).

Another **t**-test was carried out to compare the means of those who had one or more courses in education and psychological measurements. The mean of the first group was 73.34 and that of the second group was 59.83. The difference between the two means was significant at the $p < .001$ (see Table 7).

8.2. THE RESULTS OF THE ANALYSIS OF VARIANCE

The groups compared in the analysis of variance are language teachers (1), social science teachers (2), mathematics teachers (3), and teachers of sciences and other subjects (4). The means of these groups were 69.95, 66.37, 73.30 and 63.31 respectively. The results of the analysis of variance are as shown in Table 8.

The difference between the highest and the lowest means is not significant at the $p < .05$ level. So, there are no significant differences between the groups with regard to the subject they have taught.

TABLE 6

T-test between the means of male and female subjects

		POOLED VARIANCE ESTIMATE	
Mean	T Value	Degrees of Freedom	2-tail Prob.
69.886	1.18	60	.244
		SEPARATE VARIANCE ESTIMATE	
65.296	1.20	59.34	.234

TABLE 7

T-test between the means of subjects who have claimed to have one or more courses in educational and psychological measurements and those who have none

		POOLED VARIANCE ESTIMATE	
Mean	T Value	Degrees of Freedom	2-tail Prob.
73.24	3.65	59	.001
		SEPARATE VARIANCE ESTIMATE	
59.83	3.84	56.66	.000

TABLE 8

Analysis of variance of the means of groups according to the subject matter taught

SOURCE	df	SS	MS	FR	F. Prob.
Between groups	3	690.00	230.00	.9694	.414
Within groups	57	13518.24	237.16		
TOTAL	60	14207.93			

9. ANALYSIS OF THE SECOND ADMINISTRATION OF THE ATTITUDE SCALE

As mentioned before, a revised attitude scale of 21 items has been created from the attitude scale. The criteria on which the items in the new scale were selected had been discussed earlier in the chapter. The new scale was then administered to another group of teachers (71).

The main purpose of readministering the attitude questionnaire to a second sample of subjects was to examine the extent to which the dimensions extracted from the first administration of the scale were true dimensions underlying the responses of most teachers and other professionals.

9.1. INSPECTION OF PEARSON'S CORRELATION COEFFICIENTS BETWEEN EACH SEVEN ITEMS OF EACH FACTOR

To obtain such evidence, each 7 items belonging to one factor were correlated. Pearson's Correlation Coefficients have been computed between each items, to study the degree to which items in each factor were homogeneous. The significant levels of the correlated coefficients between each two items were also printed.

The results of the analysis show that all (100%) the 49 correlating coefficients between the 7 items of the first factor were significant at $P < .001$ or beyond. Of the 49 correlation coefficients between the items of the second factor, 39 (80%) of them were significant, 28 (57%) of the correlations being significant at the $P < .001$ or beyond. Of the 49 correlation coefficients between the items of the third factor, 45 (92%) of them were significant, 37 (76%) of the correlation coefficients being significant at the $P < .001$ or beyond.

However, as in the previous study, factor 1 (test accuracy) still remains a general factor on which items belonging to other factors also have substantial correlations. When an accuracy factor exists, the other two factors are implicated. Although the between factor correlation coefficients are not as high as the within factor correlation coefficients, in many cases items belonging to the second and the third factors have significant correlations with items belonging to the first factor. Consequently, there were substantial correlations between items belonging to the second and the third factors.

TABLE 9

Summary of the correlation between items in each factor

	No. of items	No. of correlations	Not Significant	Significant	Significant at .001
FACTOR 1	7	49	0	49	47
FACTOR 2	7	49	8	41	28
FACTOR 3	7	49	4	45	37

TABLE 10

The table illustrates items which have significant correlation coefficients and items which do not have significant correlation coefficients

	1	2	3	4	5	6	7		8	9	10	11	12	13	14
1	S	S	S	S	S	S	S	8	S	N	S	S	S	S	S
2	S	S	S	S	S	S	S	9	N	S	N	N	S	N	S
3	S	S	S	S	S	S	S	10	S	N	S	S	S	S	S
4	S	S	S	S	S	S	S	11	S	N	S	S	S	S	N
5	S	S	S	S	S	S	S	12	S	S	N	S	S	S	S
6	S	S	S	S	S	S	S	13	S	N	S	S	S	S	S
7	S	S	S	S	S	S	S	14	S	S	S	N	S	S	S

	15	16	17	18	19	20	21		1	2	3	4	5	6	7
15	S	S	S	S	S	S	S	8	S	S	S	N	S	S	S
16	S	S	S	S	S	S	S	9	N	N	N	N	S	N	N
17	S	S	S	S	S	S	S	10	S	S	S	N	N	S	S
18	S	S	S	S	S	S	S	11	S	S	S	S	N	S	S
19	S	S	S	S	S	N	N	12	S	S	S	S	N	N	N
20	S	S	S	S	N	S	S	13	N	S	N	S	S	S	S
21	S	S	S	S	N	S	S	14	N	S	S	S	S	S	S

	1	2	3	4	5	6	7		8	9	10	11	12	13	14
15	S	S	S	N	S	S	S	15	N	S	S	N	S	N	S
16	N	N	N	N	S	N	N	16	N	N	N	S	S	N	N
17	S	S	S	N	N	S	S	17	N	N	N	N	S	N	S
18	N	S	S	S	N	S	S	18	N	S	S	S	S	S	N
19	S	S	S	S	N	N	N	19	N	N	S	S	S	N	S
20	N	S	N	S	S	S	S	20	N	N	N	N	N	S	N
21	N	S	S	S	S	S	S	21	S	N	S	N	N	S	S

9.2. RESULTS OF THE ANALYSIS OF VARIANCE

Three different analyses of variances have been employed to examine whether or not teachers' responses to the attitude scale differed according to the subject matter taught, degree of familiarity with the construction and/or the administering of psychometric tests.

The first analysis of variance was carried out to test the differences between the means of groups of teachers teaching different subjects. As in the previous study, the differences between the means of these groups were not significant according to the subject taught. The results of the analysis are shown below:

TABLE 11

Analysis of variance of 5 groups of teachers according to the subject matter they have taught

SOURCE	D.F	SUM OF SQUARES	MEAN SQUARES	F. RATIO	F. PROB.
Between Groups	4	887.9188	221.9797	2.3991	.0592
Within Groups	64	5921.7333	92.5271		
TOTAL	68	6809.6522			

Another analysis of variance was carried out to test whether or not teachers differed according to their degree of familiarity with the construction of standardized tests. The three groups of teachers compared were:

(1) those who claimed to have been very familiar with the construction of standardized tests;

(2) those who have been fairly familiar with the construction of standardized tests; and

(3) those who were not at all familiar with the construction of standardized tests.

As shown in the following table, the differences between the means (70.80, 62.93 and 59.22 respectively) of these groups were significant at $P<.008$. The results show that the more familiar the subjects were with the construction of standardized tests, the more favourable were the responses which they have expressed toward psychometric tests, testing and test construction.

TABLE 12

Analysis of variance of 3 groups of teachers according to their degree of familiarity with psychometric test construction

SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARES	F. RATIO	F. PROB.
Between Groups	2	962.6668	463.3334	5.1980	.0080
Within Groups	66	5882.9854	89.1361		
TOTAL	68	6809.6522			

A third analysis of variance was carried out to test whether or not teachers' responses to the scale differed according to their degree of familiarity with administering psychometric tests. As shown in the table below, there were significant differences between the means of the 3 groups of teachers in the analysis. As above, these groups of teachers were:

(1) those who claimed a greater degree of familiarity with administering psychometric tests;

(2) those who had a fair degree of familiarity with administering psychometric tests; and,

(3) those who were not at all familiar with administering psychometric tests.

TABLE 13

Analysis of variance of 3 groups of teachers according to their degree of familiarity with administering psychometric tests

SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARES	F. RATIO	F. PROB.
Between Groups	2	1036.7316	518.3658	5.9263	.0043
Within Groups	66	5772.9206	87.4685		
TOTAL	68	6809.6522			

9.3. RESULTS OF THE T-TESTS

Some of the background variables of our subjects which have been investigated both in the pilot and in the main studies included sex of the teachers, the number of years he/she has been in the teaching profession, and the size of the classes he/she has usually taught. A series of t-tests were carried out to check whether or not the responses of different groups of teachers differed according to the above-mentioned variables.

The first t-test was carried out to see whether or not the difference between the means of male and female subjects in the sample was significant. The means of the male and the female subjects were 64.18 and 60.09 respectively. As shown in the table below, the difference between the two means was not significant.

TABLE 14

T-test on the difference between the means of male and female groups

GROUPS	Mean	T Value	Degrees of freedom	2-tailed probability
MALE	64.18	1.71	67	.092
FEMALE	60.09			

As shown in the following tables, a teacher's responses toward psychometric tests did not differ according to the number of years he/she has been in the profession; the number of years he/she has been teaching the subject; or the age of the pupils taught.

TABLE 15

T-test on the difference between means of teachers of over five years teaching experience and teachers of up to five years teaching experience

	GROUPS	MEAN	T VALUE	DEGREES OF FREEDOM	2 TAILED PROBABILITY
Inexperienced Teachers	1 5	62.333	.17	69	.869
Experienced Teachers	2 5	61.915			

TABLE 16

T-test on the difference between means of teachers who have been teaching the present subject for more than five years and teachers who have been teaching the present subject for up to five years

	GROUPS	MEAN	T VALUE	DEGREES OF FREEDOM	2 TAILED PROBABILITY
Less experienced	1 5	62.48	.28	69	.769
More experienced	2 5	61.80			

TABLE 17

T-test on the difference between means of teachers according to the age of the pupils taught

	GROUPS	MEAN	T VALUE	DEGREES OF FREEDOM	2 TAILED PROBABILITY
Taught younger pupils	Age 10	56.90	1.89	69	.073
Taught older pupils	Age 10	63.00			

From the results so far, it is possible to offer an answer to the question as set. The factor analyses suggest that teachers judge psychometric as well as traditional examinations on three criteria. These criteria, not unexpectedly, overlap. Firstly, accuracy is highly desirable. Teachers expect assessment scales to be free of errors and personal biases. Secondly, they expect the test to reveal knowledge of three important areas;

(a) of subject matter

(b) of the person being tested, and

(c) of the circumstances surrounding the test.

The third dimension deals with carefulness and preparation in the construction of the test. This dimension is described as the amount of objective planning and preparation.

The usual properties of psychometric test are clearly incorporated in these three dimensions. Respect for objectivity, reliability, difficulty level and validity are all implied in these criteria. However, beyond these there is an expected need for attention to other properties. Teachers seem to be asking for more individualization in assessment. Knowledge of the individual being tested and his circumstances is given greater emphasis by the professional teacher. Psychometrically speaking, too much emphasis on fitting the assessment procedure to the individual and his circumstances goes against the need for standardization of stimuli and administration procedures.

In conclusion, it seems that the objectives of assessment as teachers see them are wider than those normally encompassed in psychometric testing. The teachers felt the properties of the psychometric tests are important but other equally important objectives should be taken into account.

At this stage of the research, it seems appropriate to ask whether professional teachers are able to judge the properties they require in a test. To determine this, the second stage of this research was planned to investigate the ability of professional teachers to judge test properties.

CHAPTER 7

TEACHERS' ESTIMATES OF TEST ITEM PROPERTIES CONSTRUCTION AND DESIGN OF THE SCALE

The main concern of this section is with teachers' abilities to judge item properties.

Every normative psychometric test item can be conceptualized as though it is a multidimensional 'model' to be built. One dimension of the model might be represented by the content relevance which is to be consideredin relation to an instructional programme or behavioural objectives thought to have been attained by the pupil. A second dimension of the model could be the difficulty of the test item which is to be considered in relation to a perceived ability of the pupil. A third dimension which the test constructor has to manipulate is the discriminating power of the test item in relation to a perceived range of ability of the pupils to be tested. That is, the psychologist constructs the test item with the expectation that only a certain proportion of the pupils will get the item right.

Not only the difficulty, discrimination, and content relevance are represented in the model, but the constructor may also incorporate other dimensions into the model. The test item may well be intended for future populations. In that case, the reliability dimension of the test item has to be considered so that the item will always function in an expected way in subsequent groups or over repeated testings. In addition to these dimensions, the test constructor could have some particular assumptions about the effectiveness of previous teaching; about the backgrounds of the pupils etc. all of which enter into his overall assessment of the test item. Some more dimensions the test constructor has to account for are discussed under the test plan. The constructor must deal with all these dimensions of the test items simultaneously. He should not construct tests with little content relevance; and which are

either too difficult or too easy etc. for the pupils.

In spite of all these multiple demands, teachers set tests for their pupils or select tests for their pupils from a bank of existing standardized or non-standardized tests. When shown a sample of test items, teachers do judge the characteristics of the test items in relation to their pupils. They talk of a test item as being too difficult, too easy, good, ambiguous, as not being taught, etc. What is to be investigated here is whether these judgments are based on some intuitive understanding of the material which can be objectively verified or whether these judgments are unstable and random. If teachers' judgments of the test item properties are based on knowledge of what they judge, how well do teachers judge these properties compared with the empirical properties of the test items? In other words, if teachers predict a degree of performance of their pupils on a particular test, how far is their prediction on performance of their pupils on that particular test better than a mere chance.

1. REVIEW OF THE LITERATURE

Difficulty, discrimination, content relevance and reliability are critical properties of test items. The main deficiency of teacher-made tests is said to arise from the inability of teachers to judge these properties (Mehrens and Lehmann, 1978). The belief that psychometrically constructed tests have better content relevance, reliability and more appropriate difficulty and discriminating indices are reported by Ausubel (1969); Anastasi (1968).

Lorge and Diamond (1953) investigated the value of extra information to 14 judges of test item difficulties. The judgments of the 14 judges

were obtained under two different conditions. On one occasion, the judges were given the empirical difficulties of 10 extra test items. This information was not provided on the second occasion. First, without this information, 7 judges were asked to judge the difficulties of 45 test items taken from an arithemetic test for junior high school pupils. Then, with this information provided, the same judges were asked to estimate the difficulties of 45 test items. With the information provided, the remaining 7 judges were asked to estimate the first 45 items. Then, without the information these judges were asked to estimate the difficulty values of the second 45 items.

The competence of the judges of the test item difficulties was measured by the degree to which the mean and the standard deviation of the estimated percentages of the students passing the items approximated the empirical mean and standard deviation of the item difficulties.

The mean estimated test difficulty was 70% for all 14 judges as compared with the empirical test difficulty of 46.5%. But when provided with the empirical difficulty of the 10 extra items, the estimated test item difficulty was 53.6% compared with the empirical mean difficulty of 46.5%. Similarly, the estimated standard deviation improved with the information. The authors concluded that information about the difficulties of sample of test items improves the judgments of teachers of test item difficulties.

Lueptow, Early and Garland (1976) attempted to validate 192 college students' ratings of examination questions. They asked students to rate the discriminating power of 45 multiple-choice test items on a five-point scale.

The competence of the students' judgments was measured by correlating

students estimates of the discriminating index of the test items with point biserial values. The point biserial values formed the objective criterion against which the validity of the students' evaluations of the test items was assessed.

Results showed that most of the correlation coefficients between the students' ratings and the point biserial values were insignificant. In fact, many of them were zero. Although all students were unable to rate the test items with any accuracy, the class who majored in the subject from which the test items were taken from rated better than other classes.

The authors concluded that students were unable to distinguish discriminating multiple-choice test items from non-discriminating test items. Their conclusion implies that knowledge of the content of instruction is a qualification for a test item rater.

However they have concluded, the original design of the study was poor. First, students were not explicitly told in relation to whom the item discriminations were to be evaluated. Secondly, the subjects had no teaching experience. Thirdly, most of the students were not familiar with the content of test items.

Ryan (1968) investigated teachers' judgments of difficulty, discrimination and content relevance of test items. His aim was to determine the extent to which judgments of three test item properties relate to teachers' overall evaluations of the test items. He wanted to find out which of these three test properties contributes more to the overall assessment of the test items.

The degree of relationship between an overall item quality and the other three properties was measured by the sizes of correlation coefficients

between the judged estimates of the overall item quality and the judged estimates of each of the other three properties of the test items.

The numbers of teachers at each grade level were: 13 at the seventh and eighth, 19 at the ninth, 10 at the tenth and 17 at the eleventh. In fact, they were 4 parallel studies.

He requested judgments for each 25 multiple-choice items from each of the mathematics achievement tests. Each teacher evaluated 25 items taken from the content he taught and performed by his class and separate judgments were made for each class as a unit. The judgments for each item are as quoted:

"How good or poor is item for determining knowledge and understanding of the instructional content presented in your class(es)?

Very poor poor fair good very good

What proportions of pupils in each class will answer the item correctly?

0% 25% 50% 75% 100%

How much better will the most proficient third of the pupils in each class do on the item compared to the least proficient third?

Same as the least	Slightly better	Somewhat better	Much better	Very much better

How appropriate or relevant is the item for the instructional materials and content presented in each class?

not at all relevant	somewhat relevant	quite relevant	very much relevant

The results of the study were the following:

(a) Teachers' judgments of the relevance of items to instructional

frequently related to teachers' judgements of overall item assessment.

(b) Correlation coefficients between judged item difficulty and the judged overall item quality; and correlation coefficients between judged item discrimination and judged overall item quality were lower than the correlation coefficients between judged item relevance and the judged overall item quality.

Conclusion: He concluded that teachers' judgments of the item relevance was the most important variable in the overall evaluation of the item quality. Teachers' judgments of difficulty and discrimination were comparatively less related to the overall assessment of the item than the content relevance of the item. The estimates of the two properties, difficulty and discrimination, were not quite independent of the content relevance of the item. Teachers could make more accurate estimates of difficulty and discrmination when the content was more familiar than when the content was less familiar.

2. APPROACHES TO ASSESSING TEST ITEM QUALITIES

There are priori and posteriori approaches to the evaluation of test item properties. The first is an expert judgment used to select suitable items and to eliminate inferior items. The second is an empirical approach which is the main source of the quantitative information about the test items. Some suggest that content experts can rate test items more quickly with a high degree of validity and reliability (Rovinelli and Hambleton, 1976). The arguments propounded by the proponents of the first approach imply that systematic human judgment of the test item qualities can be enough for the ensemble of suitable test items. Therefore, any more benefits obtained through quantitative application of

item analysis of the test items would not be commensurable with the extra labour incurred in such empirical analysis. For them, quantitative analysis of the test item properties is only supplementary to the qualitative assessment of the test item qualities. They argue that the best way of ascertaining the appropriateness of a test is an examination of the test by competent judges (Ebel, 1956; Wesman, 1971).

On the other hand, those who support the alternative approach, the quantitative item analyses, argue that human observations can never be systematic enough to account for all problems to be encountered in the process of test construction. They argue that there are always some adverse conditions under which test items are to be constructed. These adverse conditions include insufficient time, lack of adequately trained personnel etc. This means that there are too many prerequisites for the test constructor all of which cannot be fulfilled by any one person. The test constructor is required to have a thorough knowledge of the subject matter; an intimate understanding of the instructional objectives, the experiential backgrounds, abilities and the mental processes of the subjects who will take the test; facility in clear language; willingness to devote time and energy (Conrad, 1949). Due to these adversities the psychometricians argue that the subjective examination of the content to determine the representativeness of the test of the behaviours to be measured can be deceiving. The empirical approach frequently employed in the assessment of the test item qualities is the item analysis.

3. THE IMPORTANCE OF ITEM ANALYSIS

Item analysis has been defined as "a....statistical or quantitative procedure used in psychometric or test construction for determining the

suitability of any specific test item for inclusion in a particular test" (Jensen, 1980). Item analysis is important for the construction of almost all tests. The importance of item analysis is that it improves the quality of tests. (Anastasi, 1976; Lange et al., 1967). The most desired qualities of a test are its validity and reliability. Validity and reliability of a test depend very much on the characteristics of the individual test items. These qualities, as well as many others, can be built into the test, in advance, through item analysis procedures (Gronlund, 1981).

Item analysis can be utilized to provide valuable information, about examinees' responses for assessing the nature and the amount of learning or for determining the causes of learning difficulties. It reveals the areas of instructional weakness that needs more attention. Secondly, item analysis provides insight into preparing better tests for future assessments. Experience in item analysis increases one's general skills of test construction. One learns how to avoid ambiguities, clues, ineffective distracters and many other technical defects in test construction.

Other fringe benefits of item analysis are:

(1) that item analysis provides adequate basis for efficient class discussions of the test results. Knowledge of how items in a test functioned helps the teacher to save time and avoid unnecessary arguments with pupils in the class.

(2) Item analysis permits one to use shorter tests without sacrificing the validity and reliability of his test.

(3) Item analysis makes it possible to build item banks. From these item banks one can retrieve appropriate test items in accordance with one's needs.

4. Indices and Procedures of Item Analysis

There are many indices and procedures of item analysis. Only those indices and procedures of item analysis, which are more relevant to the present study will be discussed in this section. They are the psychometric properties we have asked the teachers to judge. These properties which have been investigated in this questionnaire are the difficulty of the items, discrimination power, content relevance, reliability and the overall quality of each of the 24 test items. At a later point we shall ask students, teachers and other professionals to compare two sets of tests on the basis of 15 criteria of test qualities. The purpose of the present discussion is to introduce issues involved in these indices and the procedures which are relevant to the test qualities to be estimated by teachers. We shall point out the importance of the information obtained from these indices and the consequences of failure to optimize this information.

5. Types of Information obtained from Item Indices

Before treating them individually one must look at the types of information supplied by these item indices. The most important types of information come from three main dimensions of the test. These are as follows:

(a) the item as a whole

(b) the individual choices offered by the item and

(c) persons attempting each item.

The information provided by the item as a whole includes the difficulty index, the discrimination index, measures of item-criterion correlations i.e. item-total correlation, item-external criterion correlation, etc.; number of omissions or persons who failed to record

response for an item, etc.

Information provided by the individual choices include the number of persons selecting a given choice as the answer; direction of discrimination of each alternative, etc. Information provided by the sample includes the number of persons attempting each item, mean and standard deviation of those attempting each item etc.

6. Test Item Properties

6.1. Item Difficulty

In normative testing, if the items are too easy or too difficult they provide very little information or no information at all, depending on how easy or difficult they are. Suppose a test given to a group of subjects was so easy or so difficult that all candidates either passed or failed, the examiner would not be able to obtain any information concerning the individual differences. If all candidates get either zero or a full mark, there are no differences between any two individuals, with respect to what the test items measured. The total lack of information occurs when item difficulties are either 1 or 0. Maximum information, on the other hand, is obtained when the difficulty of the item is about 50%. As one moves towards the middle the more the item differentiates and the less it differentiates as one moves towards either extreme. If 50 candidates pass an item and another 50 candidates fail the difficulty level of the item is 50% (p =.50). The information yielded by the test item is 2500 units (50x50). The figure is the number of pairs of comparisons or differentiations between each two candidates. An item passed by 80% of the candidates would yield 1600 units of information. One passed by 15% would yield about 1275 units of information. An item passed by 100% would provide 100x0=0, or no information. Similarly,

one passed by no candidate would provide 0x100=0 or no information. According to the latter two items all candidates have exactly the same ability, achievement etc. In summary, the amount of information provided by an item dwindles (Conrad, 1949; Guilford and Fruchter, 1978) as one moves toward the extremes.

Not only does the item yield more information at the middle difficulty level but items at that difficulty level become very homogeneous. They have higher intercorrelations. This means that they (items) elicit similar responses from the candidates. Homogeneity and higher interitem correlations are desirable characteristics of test items. Still another reason why 50% difficulty is preferable is that at that difficulty level the item has the highest discrimination index.

However, for other reasons, test experts do not usually recommend the selection of .50 difficulty level of all the items in a test. They suggest items of various difficulty levels, but with an average difficulty level of .50. That is, some items could be easier or harder than .50, but the average difficulties of all items in a test should be about .50.

The difficulty of the test items influences the spread of scores, or test variance. The widest distribution of test scores is obtained by items of the .50 difficulty (Anastasi, 1976;1965). The nature of the distribution of scores, in turn, influences the reliability of the test.

Still another purpose of manipulating test item difficulties is for the optimization of screening candidates at a particular selection ratio.

If the selection ratio is the upper 20% of the candidates, for example, the best items are those clustering around a p of .20 (Anastasi,

1976; Conrad, 1949).

Why do we measure the difficulty level of test items? The main reason for measuring the difficulty level of a test item is to choose a suitable difficulty level for a group of candidates. A suitable difficulty level of a test item is present when the item or the test as a whole is not too difficult or too easy. The difficulty level of test items is not a fixed property of the test item. One cannot talk about the difficulty of a test item in isolation. The difficulty of a test item can have meaning only when it is related to known population. The difficulty level fluctuates with the group ability and the degree of familiarity of the testees with the content of the item. (Scores of the re-test is a good example to illustrate this point). It also depends on which approach of testing, (criterion-referenced or norm-referenced testing) is at issue.

Secondly, item difficulty is used to ensure student motivation in test taking situations (Gronlund, 1981; Anastasi, 1976). Teachers try to control and eliminate extraneous variables, to the central purpose, which may influence the performance of the examinees on the test. They do so by arranging test items from the easiest to the most difficult. The reverse of this arrangement has been found to have a depressing effect on testees' confidence in taking the test (Anastasi, 1976).

Thirdly, the difficulty of the test item is related to the amount of information it provides. In normative testing items of medium difficulty provide more information than items which are too easy or too difficult (Anastasi, 1976). The most important information in normative testing is the amount of individual differences the item can depict.

6.2. Item Discrimination Indices

Item discrimination is a validity index. There are many test validity indices (Anastasi, 1976). According to Anastasi (1976), all the procedures of test validity yield similar results. She suggested that the best method is the one that requires the least computation. Item validity indices are based on the relationship between examinees responses to an item and the total test score (Wilmut, 1975). This relationship in turn is based on the assumption that a student with a high criterion score has a higher tendency (probability) of choosing the right answer to any item than a student with lower criterion score (Henryson, 1971). So, item discrimination is defined as the ability of the test item to differentiate between students of high achievement and students of low achievement. However, the direction of differentiation is essential. An item could be positively discriminating, negatively discriminating or non-discriminating. The item is positively discriminating to the extent to which the upper group scores proportionally higher on the item than the lower group. The item is negatively discriminating to the extent to which the lower group scores proportionally higher on the item than the upper. The extent to which the upper and lower groups obtain equal scores, indicates that the item is non-discriminating. Only positive discrimination is considered good. Positively discriminating items contribute to the positive functioning of the test. These items add something to whatever the test is measuring. In other words, more able students have a higher probability of providing correct answers to any test item. According to this assumption, items which fail to agree to the above-mentioned relationship are declared inferior.

Discrimination indices employ internal criteria. Two measures of the item validity commonly used are the biserial and point-biserial

choice depending on certain assumptions about the score distributions (Guildford and Fruchter, 1978; Henryson, 1971). However, both procedures are difficult in computation.

Another disadvantage of these procedures is that they are applicable only under certain assumptions. Therefore, to make sure that the desired distribution is present, one has to check, first, the nature of the score distributions.

Point-biserial procedure imposes fewer assumptions to be held true about the score distributions. The biserial procedure demands that the criterion scores for all examinees to have normal distribution and that the actual scores on item, although dichotomized (Scored 0,1), to be normally distributed. The point-biserial procedure, on the other hand, assumes that the dichotomized variable (item) has a true dichotomy. (Male, Female). Another argument for the point-biserial is that it tells more about the contribution of an item to the validity of the total test than biserial does (Guilford, 1965). If the scores on the criterion are not normally distributed, the biserial correlation coefficient will be higher than 1.00. This situation contradicts Pearson's correlation coefficient of which the biserial and point-biserial are variants.

One more assumption made is that the test as a whole is measuring only one thing. If that assumption does not hold true and the test is measuring different traits, item-total procedures would not have any meaning. When the scale is not unidimensional, items should not be expected to show high item-total correlations.

Thirdly, all item-total correlations have spurious correlation coefficients specially when the number of the test items are small. Item discrimination index is related to the test reliability. That is, the

higher the discrimination index the higher the reliability. As discussed in an earlier section, item discrimination is also related to the difficulty values. It is biased in favour of an intermediate item difficulty. That is, as the difficulty level approaches .50 the higher the discrimination index (Anastasi, 1976; Blood and Budd, 1972). However, item discrimination is not always dependent on the difficulty level. When discrimination index independent of the difficulty is required, one should use the biserial procedure.

A third commonly used procedure of item discrimination index is the upper-lower criterion groups. These groups usually involve the upper 27 and the lower 27% of the candidates (Kelly, 1939). However, other percentages may be used depending on the shape of the score distribution. The upper and lower percentages are desired to be greater than 27% when the distribution is flatter than the normal curve (Cureton, 1957).

6.3. Item Content Relevance

Content relevance is an aspect of validity which describes the item-objective congruence. The other aspect of the validity is reliability. By content relevance or content validity of the test item, we mean whether or not the test item is measuring the educational objective taught in the class. Here, content refers to the subjects' universe of behaviours. What we asked the teachers was whether or not the behaviours called for by the test items were those which have been taught in their classes. We did not intend to ask them about the processes by which each of these behaviours would have been elicited.

The educational objectives are the desired behaviours expected to have been acquired by the candidates (see next section). These include

recall and recognition of knowledges; understanding; applying; organizing and evaluating ideas (Bloom, 1956). The relevance of the item corresponds to the test items' representativeness of these objectives which has been stated in behavioural terms.

The degree of relevance of the test item is determined by empirical and/or judgmental procedure(s). In the judgmental approach, one can rely on the available professional judgments to establish the item's relevance to specified objectives. In the first approach, one assumes first that the whole test is measuring the trait it was intended to measure. With that assumption, then one correlates the item with the total test. If the item has significant correlation with the total test, one has confidence that the item is measuring the trait which the total test was measuring. If it is so, then the item has content relevance. However, that assumption could be false and the test could be measuring different things or a wrong thing. The legitimacy of making assumptions about what particular tests measure is discussed below.

6.3.1. Measurement of Content Relevance of Test Items

Content relevance indicates the extent to which the content of the test reflects the content of the behaviour or property being measured. This can be done directly or indirectly. A direct measure of content relevance is obtained in achievement tests when the items of the tests adequately cover the body of knowledge taught. An indirect measure is deduced when an empirical link is established between the property and the test item. Such indirect links are commonly used in aptitude as distinct from attainment tests. The use of reaction time to indicate driving potential is only content valid when a reliable correlation has

been demonstrated between reaction time and driving ability. Similarly, the use of an individual's reaction to an inkblot can only be used to indicate personality characteristics if a reliable correlation exists between them both. In other words, content relevance of the test item is the match between the item and the objective to be measured by the item.

The most important question to be dealt with in content relevance of an item in a test is the way in which one can know the extent to which a particular item measures whatever it was intended to measure. There is no direct measurement in testing. All the evidence available to us is based on inferences (Popham, 1980). Therefore, our knowledge of the extent to which test items measure particular content is true only to the extent to which the inference itself happens to be true.

One common inference in relation to item relevance is made when the item is investigated to find out whether or not the item is measuring the same thing that other items in the same test, as a group, are measuring. There are methods to demonstrate that a test item is, or is not, measuring the same thing that other items in the same test, or even an item in other tests, are measuring.

Weaker than making inferences about what the item measures is the fact that our knowledge of what the test itself measures is based on assumptions. Most empirical procedures on item validity assume the test, or criterion, to be measuring the right thing. When the assumption does not hold true, the inference made that the item has content relevance or is measuring the right trait becomes false.

The basic problem of item relevance does not pertain to whether or not one can determine the relationship between an item and the rest of the test, the criterion, but mainly pertains to whether or not one can

show that the test itself is measuring the right thing. If it can be shown that the test is measuring the right thing it can also be shown that the item is measuring the right.

Knowing whether or not the test measures the right thing is a question of test validity. But one thing that would be clear in all the literature is that, we do not depend only on blind assumptions for our knowledge of what tests measure. The next section discusses reasons for making assumptions about what tests measure.

6.3.2. Justifications for Making Assumptions about what Tests Measure

Why do we make assumptions about what a particular test measures? There are several reasons that make one assume that a particular test does or does not measure what it was intended to measure. The reasons for making such assumptions are many and varied. First, of what is generally known about testing, one assumes that the plan of the content and the construction of the test items have been carried out properly. Second is the test's face validity, or the apparent match between test items and the behavioural content of the test. Third is a more systematic human judgment. Fourth is concrete empirical evidence obtained for the test.

In most cases we take it for granted that the plan of the test content and the item construction are carried out systematically, or if not systematically, at least they are not randomly executed. One believes that there is every reason why test constructors should always try their best to select test content and construct test items systematically. These assumptions and efforts are convincing reasons that one's assumption of what the test measures is justifiable.

The assumption is valid because it is based on the known fact that

most tests are, more or less, systematically constructed and not randomly assembled; that the domain of the test can be meaningfully defined and adequately sampled (Lennon, 1956).

The second evidence that takes one to make assumptions about what a particular test measures comes from a simple inspection of the test items, usually called the test's face validity. Face validity of the test is the most rudimentary source of evidence that a particular test measures what it was intended to measure. However imprecise are decisions based on face validity, the simple appearance of the test narrows the range of possible domains that a test can measure. For example, no matter how poor a judge one might be, it is unlikely that he would agree that a test intended to measure arithmetic skills would measure knowledge of English grammar. But one should remember that when the domains in question become more similar, decisions based on the test's face validity become less valuable. However, the face validity of the test is enough to give the test evaluator a great deal of confidence to make assumptions that the test is measuring a particular trait (Nunnally, 1970; Flanagan, 1939; Burroughs, 1975).

A third evidence that takes one to make assumptions about what test measures comes from systematic human judgment of the match between test items and the behavioural objectives. The systematic human judgment of the test validity begins with the ways in which the domains of behaviour which the test to be generated would be measuring have been defined. The second phase concerns the generation of test items that measure these behaviours. Thirdly, the test is subjected to judgmental validation. The test is then said to be valid to the extent to which judges think that test items apparently match the behavioural content, or the test has an item-objective congruence (Berk, 1980). Between these

three phases there can be many intermediate steps of test item validation.

There are several endeavours made towards achieving systematic human judgmental validation of tests which are discussed elsewhere in this study. The reader is referred particularly to the works of Ebel (1962a); Popham and colleagues of the Iox; Hively, Patterson & Page (1968); Hambleton (1980).

Some test experts contend that systematic human judgment of the test is enough to determine whether or not a particular test is measuring what it was intended to measure (Ebel, 1962; Nunnally, 1972; Thorndike & Hagen, 1977). Others do not believe that human judgment alone is enough to determine test validity. They would rather be confident to make assumptions about what tests measure only when there is an empirical evidence obtained from examinees' responses (Messick, 1975; Linn, 1979). The controversy is over whether the validity of the test inheres in the test items as stimuli or in the examinees' responses made on the test items (Hambleton, 1980; Lennon, 1956).

The final source of evidence that permits one to make assumptions about what tests measure comes from a variety of empirical procedures. First,the test in question is correlated with other tests known to have been measuring the same trait (Burroughs, 1975). If the test has positive correlations with these tests the test is said to be valid for the purpose it was intended to achieve. Secondly, if the internal correlations among items in the test are high enough, the test is said to have validity. The intern al correlations among items in a test would be low when items in the test are measuring different traits (Anastasi, 1976). This procedure can be criticized on the grounds that all the items in a test could be measuring the same wrong thing and hence can have high correlations with one another. But the probability that all items in

a test are measuring the same wrong trait is very remote. Third, the comparison of performance on the test before and after a period of training is considered to indicate the validity of the test. That is, if the trait measured is susceptible to training, scores on the test are expected to increase from before to after. This type of validation is called instructional sensitivity (Berk, 1980). Fourth, other evidence that support assumption-making about what tests measure come from construct validation of the examinees' responses to the test in question Proponents of empirical procedures argue that judgment of the test's content validity is not sufficient because it does not address itself to the scores made on the test. Since descriptions and decisions are made on the scores, not on the test items, test validity must address itself to the scores (Hambleton, 1980).

Since an assumption made about what a particular test measures is based on evidence, the assumption is said to be a reasonable one (Ebel, 1979). Ebel added that "item analysis using an internal criterion (total score) makes a test a better measure of whatever it does measure" (p. 261). Ryan (1951) considered internal criterion to be a better measure of item validity than external criterion. But, Ma scuilo and Slaughter (1981) expressed serious concern about the validity of using internal criterion in assessing item validity.

6.4. Item Reliability

In the item reliability, we were interested in how much teachers thought that retest scores of a test would be influenced by a pretest of the examinees.

The procedure required two administrations of the same test to the same group of subjects and the correlation of the two sets of scores.

If there is a short interval between the two testing occasions, memory has spurious effects on the results (Stanley, 1971). There is an over-estimation of the reliability coefficients of the test in question (Anastasi, 1976). Reliability estimates vary with the length of time that elapses between the two testings. When interpreting test-retest reliability coefficient one has to consider the length of the time interval.

To minimize memory effects, one must allow a longer interval to elapse between the two administrations of the test. The longer the interval, the less the effect of memory becomes. The amount of memory effect depends on the length of the test, and the nature of test content. The shorter the test, the greater the effect of memory. Longer tests are less susceptible to memory effect. Other things being equal, more distinctive test items are better remembered.

However, the disadvantage of the extended interval is that the trait measured by the test changes, unless it is assumed to be stable over time. (Burroughs, 1975). The source of variance is not only due to changes in time interval, but an unknown amount of learning; pupils' differences in capacity and the amount of interference etc. Other sources of fluctuations include different levels of motivation in taking the test; mood of the testee; his health status etc. (Stanley, 1971; Gronlund, 1981). In short, it is difficult to eliminate all error variances. In this study, 7 days has been chosen to be the time interval allowed to elapse between the first and the second testings.

Our task is to obtain the teachers' judgment of these psychometric properties of the test items and correlate them with calculated or empirical values of these properties.

Secondly, we want to find out how each of the 4 properties mentioned above enters into the overall assessment of test item quality. That is, which of these properties is more related, or contributes more, to the overall item quality.

The degree of relationship found between teachers' ratings of the psychometric properties and the empirical psychometric properties of the test items is taken to indicate the adequacy of teachers' estimates of the item properties.

To find out how teachers' estimates of difficulty, discrimination content relevance and reliability are related to the teachers' overall assessment of the test item quality, each value of these 4 properties will be correlated with the values obtained for the overall quality of the test item. The degree of relationship between the overall item quality and each property will be taken to indicate the relative importance of each property to the overall assessment of the test item quality.

The purpose of the investigation is to obtain a better idea about the teachers' competence to estimate the psychometric properties of the test items. The results of the experiment are intended to provide better information to study and understand the relative efficiencies of the psychometric and the professional procedures of testing and test construction.

It will be concluded that teachers are able to estimate test item properties adequately or that teachers are not able to estimate test item properties adequately. The criterion of adequacy is determined by how well teachers' rating scales achieve its prediction, compared to the empirical values. If it is concluded that teachers are able to estimate test item properties adequately and there are significant correlations

between teachers' predictions of the performance of their pupils and the actual performance of the pupils, then, psychometric procedures of test construction could be seen as redundant and extra labour. On the other hand, if it is concluded that teachers are not able to estimate test item properties adequately, the psychometric procedures of test construction are necessary to improve (if they do so) subjective procedures of test construction and substantiate human judgment.

The subjectively constructed tests are valid and reliable to the extent to which teachers are able to estimate the test item properties.

7. METHOD

7.1. Subjects

The subjects were 22 teachers teaching the content from which the test items to be evaluated were taken from and their classes. It was assumed that these teachers were fairly familiar with the content of the subject matter they teach, as generally expected.

7.2. MATERIAL

The material consisted of 24 test items, all multiple-choice type items drawn from the content that has been taught by the teachers who rated the test item properties. The multiple-choice items were chosen for reasons outlined by Nunnally (1978).

7.3. PROCEDURE

Teachers were asked to rate 5 aspects of each test item. These were the difficulty, discrimination, reliability, content relevance and

the overall quality of each test item. Judgments of the teachers for each item property were asked for on a five point scale. Each of the five points on the scale were further subdivided and assigned numerical values. The numerical values assigned to the 10 response categories increased with the verbal ratings from lowest to highest. The scale allows standard procedures of responding to the stimulus and recording the data for all the items. The quantification of the categories also permits a straightforward computation of the data. Also, the use of technical terms in the scale were avoided.

For determining data for the item reliability, the test was readministered to the same subjects after 7 days. The actual mode and the instructions are shown on page . The text of the scale is in Appendix I.

7.4. ANALYSIS

First, from the pupils' responses the empirical difficulty, discrimination and the reliability indices and the overall quality of each item were obtained. These empirical values were correlated with the teachers' estimates of the item properties. The empirical values obtained from the pupils' responses to the test items formed the criterion. Teachers' competence to estimate the test item properties was measured by the degree of correlation between these two sets of values.

8. THE SCALE FOR TEACHERS' JUDGMENTS OF TEST ITEM PROPERTIES

8.1. Instructions

For each test item presented on the following pages we ask you to make judgments concerning certain properties of the items. The five test item properties to be judged are the overall item quality, the difficulty of the item, the discriminating power of the item, the content relevance

of the item, and the reliability of the item. Each item to be rated is followed by five different questions, each asking for a different item property to be judged.

I would like you to judge each item on the relevant point on the scale. Please tick the empty box on the scale which corresponds to your judgment. If, for example, you think the item is very good, you put your tick in the box under the response category, very good. In Questions A,C and D you will have a further two choices under each response category to the left and the right of the line which subdivides the space under each response category.

THANK YOU

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor		Poor		Fair		Good		Very good	

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same		Slightly better		Somewhat better		Much better		Very much better	

D. How relevant is this examination question to the material you taught in class?

Not at all relevant		Somewhat relevant		Relevant		Quite relevant		Very much relevant	

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

To examine whether teachers' estimates of the overall quality of the item has objective bases the following steps were taken:

(1) all test items were rank ordered according to teachers' estimates

(2) all test items were rank ordered according to psychometric properties; and

(3) the teacher estimates of the test item quality and psychometric estimates of the test item quality were correlated.

To obtain teachers' ratings for item difficulty, the teachers were asked to say how many of the pupils in their classes would pass a particular item; to obtain teachers' ratings for the item discrimination, the teachers were asked to say how much better the top third of their classes would perform the item compared to the bottom third of their classes; to obtain teachers' ratings for the item validity, teachers were asked to estimate the extent to which the item in question was relevant to what has been taught; and so far the reliability of the item.

Accuracy of the judgments being something else, these judgments had solid legitimate bases to stand on. Teachers can always judge the overall quality of the test item. But, unlike other indices, there is no overall value of the test item directly obtainable from student performance. A criterion for estimating an overall item quality from student performance is discussed in the following section.

8.2. CRITERION FOR STATING OVERALL ITEM QUALITY

To rank order test items according to their overall quality, one must state the criterion to be used. There exist criteria for rank ordering test items according to some individual characteristics such as difficulty, discrimination, reliability etc. Obviously, some test item properties have been mentioned more often than others to carry more weight for the overall assessment of the quality of the test items. But there is no agreed criterion or set of criteria, on which items in a test can be rank ordered according to their overall quality.

For norm-referenced achievement tests, the overall quality of a test item can be best determined by one of four psychometric indices. According to Nunnally (1970), the foremost important psychometric index first to be considered for the overall assessment of the test item is the item's relationship with the total score of the test. Flagnan (1939) suggested that the "best index of item validity is one which provides an index of the extent to which an item will predict the criterion" (p. 677). Guilford and Fruchter (1978) contended that item-total relationship is more important than item difficulty for the overall assessment of test items. Jensen (1980) supported the item-total relationship to be the first criterion to be considered in item selection.

The correlation coefficient computed between the item and total test score must be high if the quality of the test item under investigation is to be rated good. Items which have high correlations with the total score are the most valid items. They are most likely to be less ambiguous, to have appropriate difficulty level, and have much to do with what the rest of the test items are measuring (Nunnally, 1970). It provides more information about the test's construct validity as well. When

item-total procedure is employed the psychological variable(s) measured become more clear and uniform for all the items or items share one common factor. But the main reason that item-total relationship procedures have been stressed most is that they tend to favour those items in the test which have other desirable psychometric indices for the overall assessment of the item quality.

The measure of item validity proposed is the biserial correlation coefficient. There are two reasons for proposing biserial correlation coefficients to be computed as a measure of item validity. Biserial correlation yields a measure of item-criterion relationship independent of item difficulty (Anastasi, 1976). Although units of analysis were rather small, particularly for a biserial procedure, the present data reasonably satisfied the continuous and the normal distribution assumptions underlying the use of the procedure. The data also permit the artificial dichotomization of the otherwise inherently continuous scores of the item (Guilford and Fruchter, 1978; Nunnally, 1970). This has been indicated by the sizes of the correlation coefficients between the total test score and score on the item. There were no correlation coefficients exceeding 1.00 except on one or two occasions.

Higher distrimination index is another desirable psychometric index of test items. All items in the test must discriminate well between examinees. Test items which discriminate best between individuals are those which have middle P values (near .5). Item-total correlation procedures tend to favour items of middle difficulty level (Nunnally, 1970; Popham, 1981).

Item discrimination is a useful index of test item quality. Pyrczak (1973) investigated the validity of using item discrimination index to detect the presence or absence of faults in test items. He compared

three tests of 3 levels of faults: No faults, moderate extent of faults, and severe extent of faults. Pyrczak concluded that item discrimination is a valid measure of item quality. He also concluded that item discrimination procedure cannot identify all the items with faults. This is because item discrimination indices are influenced by factors other than faults in test items, such as the difficulty level of the item.

Item-total correlation reflects the relationship between examinees' performance on a criterion (total score) and their performance on a particular item in question. The higher and more positive the item discrimination index the better the item is (Popham, 1981; Ebel, 1979). Item-total correlation procedures tend to favour items of higher discrimination index.

A third desirable psychometric index which the item-total relationship procedures tend to favour is the reliability of the test. The higher the item-total correlation coefficients of items with the test are the higher the reliability (α or K20) of the test would be (Guildford & Fruchter, 1978). Item-total relationship influences the test reliability via item P values.

The final criterion for the overall assessment of the item quality is the difficulty level of the test item. Item difficulty is an important criterion of item selection. The difficult level restricts the degree to which items in a test can correlate with one another and can discriminate between examinees (Popham, 1981). Items with different p values tend to have low correlations with one another; those with similar p values tend to have higher correlations with one another. However, with extreme p values in either direction, items tend to have low correlations with one another.

items of extreme p values do not discriminate well between examinees, they have small variances (pq) which in turn results in the low reliability of the test.

On the other hand, items of middle p values tend to have higher correlations with one another. They discriminate well between individuals and tend to have larger variances and higher reliability indices (Gronlund, 1981).

To rank the test items according to their overall quality the following criteria have been adopted: for the item-total relationship, the discrimination, and the stability of the item over repeated testings, the higher and more positive the validity and the reliability indices are, the better the item would be. For the difficulty level, other things being equal, the closer the difficulty index to .5 the better the item would be. For rank ordering items, we have adopted item-total relationship to be the criterion on which items can be ordered according to their overall quality.

CHAPTER 8

ANALYSIS AND THE RESULTS OF TEACHERS' JUDGMENTS OF ITEM PROPERTIES

Twenty two teachers were asked to rate 6 psychometric properties of each of 24 multiple choice test items. The teachers were provided with a scale of ten categories constructed for rating the item properties from lowest to highest. The item properties rated were the overall quality, the difficulty level, the discriminating power, reliability, and validity of each of the 24 items.

As shown on page 130 the scale was presented to the teachers in a non-technical and more meaningful language. For example, instead of asking the teachers the difficulty level of a particular item, the teachers were asked to judge the percentage of their classes they think would pass a particular item. All judgments on item properties were made in relation to the teacher's own class of pupils. Where teachers have more than one class, they were required to indicate the particular class in relation to which they have judged the test items. Only these classes performed the 24-item test.

The 24-item test was then administered to students in 22 classes taught by the teachers who rated the item properties. The scores made by the students, in each class, on the items were psychometrically analyzed and a set of values for the above properties were obtained for each of the 24 items.

The two sets of values for each item property are paired as shown in Tables 1,2,3,4,5 and 6. For each item property, A,B,C,D,E and F there are two sets of values over the 22 classes. One set of values came from the teachers' ratings of the 24 test items and the other has been estimated from the students' performance. Both sets of values are

expressed in percentages. In each class, the rows preceded by the 'T' are the teachers' ratings of each of the 24 test items. The rows preceded by the 'S' are the values estimated from the students' performance.

As the scores stand in Tables 1,2,3,4,5 and 6, the degree of relationship between teacher ratings and class performance is not very clear and therefore not easy to interpret. To reduce the data and obtain a meaningful summary, each two sets of values for each of the 22 classes should be correlated. The procedure used was Pearson Product-Moment coefficient of correlation. The results of the correlational analysis are discussed below.

As suggested above, the teacher's ratings of the item properties and the psychometrically obtained values for the same item properties were correlated. Before determining the degree of the relationship between the two sets of values, the combined ratings of the 22 teachers for each item and the combined values computed from the performance of the 451 students on each item were obtained. The correlation coefficients of the two sets of values are depicted in Table 7.

1. THE RESULTS

The results shown in Table 7 are correlation coefficients between teacher rating of six test item properties of each of 24 test items and scores of psychometric indices obtained for the same item properties from students' performance on the same 24 test items. Twenty two teachers rated the item properties and 451 of their pupils performed the test.

As shown in the table, about half of the correlation coefficients

TABLE 1

The overall quality of each of 24 test items as rated by teachers and as estimated from students' performance on the test items

CODE	CLASS		TEST ITEMS 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
101	COL 1	T	70	60	70	60	70	80	90	-	70	70	90	70	70	90	90	70	70	90	50	80	90	50	70	70
		S	00	36	44	73	78	68	00	76	59	22	93	93	48	75	56	20	59	25	10	44	00	00	93	C0
102	COL 2	T	90	80	40	90	90	80	90	80	90	90	80	80	80	90	90	50	90	90	40	60	90	60	60	90
		S	63	56	51	00	50	23	00	80	19	35	00	27	38	67	42	00	00	63	26	00	80	31	31	23
103	COL 3	T	70	60	50	60	50	70	80	60	30	60	80	50	40	80	70	50	60	50	50	50	70	60	50	70
		S	70	60	10	10	10	10	80	40	50	60	30	70	20	80	60	30	30	20	30	30	40	20	20	50
104	COL 4	T	80	70	70	80	80	60	40	50	60	80	70	60	70	80	60	50	50	50	40	50	40	40	30	70
		S	00	19	72	40	21	00	00	17	69	71	41	15	00	23	73	00	21	14	00	28	24	00	20	41
105	COL 5	T	60	80	40	60	70	60	60	60	60	60	50	50	60	60	60	60	50	60	50	30	60	40	60	40
		S	00	81	28	92	03	53	00	46	39	42	76	31	64	28	50	34	06	09	15	46	00	00	44	03
106	COL 6	T	80	80	80	60	80	60	60	40	90	80	80	70	60	80	60	90	90	80	70	40	80	60	80	90
		S	00	00	39	87	35	40	00	43	45	56	38	00	20	39	35	00	27	73	31	10	00	43	18	58
107	COL 7	T	70	90	90	70	50	70	90	70	90	80	80	80	40	60	60	80	100	80	40	60	80	60	100	70
		S	100	89	55	53	02	00	48	45	37	84	05	24	19	80	16	44	20	21	00	20	83	08	33	53
108	COL 8	T	80	30	10	10	10	10	80	60	80	90	90	60	30	40	10	10	40	10	20	10	10	20	20	90
		S	87	71	67	70	71	81	75	56	73	91	59	00	00	27	38	63	56	76	00	17	73	36	00	77
109	COL 9	T	100	80	100	100	100	80	80	60	10	70	80	60	60	50	70	60	50	50	50	50	70	60	60	60
		S	12	83	19	93	28	59	00	37	34	23	73	28	59	31	39	44	16	17	13	52	00	00	39	00
110	COL 10	T	50	50	70	50	70	50	70	30	70	70	70	50	50	70	70	70	50	70	50	70	70	50	50	70
		S	00	31	58	72	10	49	00	29	37	58	90	00	00	59	51	53	74	43	00	94	89	78	30	69
201	STH 1	T	10	30	50	10	20	30	10	10	10	10	30	10	50	10	10	50	-	50	40	50	10	10	30	10
		S	19	55	41	59	24	23	00	19	39	81	82	14	32	38	41	29	53	50	31	00	00	22	47	11
202	STH 2	T	70	60	60	70	80	70	70	60	50	70	80	70	70	60	70	90	50	70	50	30	70	-	80	80
		S	00	67	04	42	50	39	00	05	69	86	59	30	40	14	43	00	38	83	31	00	00	21	47	44
203	STH 3	T	80	50	90	60	90	90	60	80	100	80	80	90	80	90	80	90	-	80	90	90	90	80	90	80
		S	31	61	21	58	60	61	00	15	81	79	54	00	44	12	50	14	41	87	00	00	10	13	46	35
204	STH 4	T	70	90	90	90	90	90	90	90	90	90	90	90	90	90	90	70	90	90	50	70	90	90	90	90
		S	52	67	83	70	00	48	00	48	26	89	51	32	13	65	75	44	44	44	47	18	63	00	47	51
205	STH 5	T	60	-	80	60	50	60	100	90	70	60	50	50	90	80	70	60	-	100	100	40	100	50	60	100
		S	54	67	80	72	00	50	00	46	28	83	48	30	13	51	76	42	41	46	41	16	70	00	43	54
2D6	STH 6	T	70	70	70	70	70	70	70	10	50	30	70	50	10	70	70	50	50	70	70	30	70	10	50	60
		S	43	63	53	29	18	80	00	63	32	76	100	50	45	30	49	57	23	00	77	01	48	00	48	69
301	MAN 1	T	70	70	30	90	60	80	80	60	80	60	80	30	80	80	70	50	50	70	70	30	70	10	50	60
		S	67	27	26	44	81	66	00	51	43	97	71	00	00	01	56	22	18	83	00	24	34	17	00	82
302	MAN 2	T	70	80	90	70	80	30	100	50	90	50	70	50	60	80	70	60	10	83	80	40	80	20	50	80
		S	51	51	34	00	70	26	00	50	49	57	34	06	35	61	01	03	65	100	04	45	45	59	33	37
401	BUC 1	T	50	80	40	80	70	80	80	60	50	80	70	70	40	60	70	70	50	70	70	20	30	30	40	60
		S	00	37	01	77	07	61	48	96	49	84	83	53	10	37	56	54	50	18	56	32	53	16	63	82
402	BUC 2	T	10	50	70	70	90	70	70	90	70	50	90	50	70	90	70	90	60	50	90	50	70	50	90	70
		S	00	00	29	32	100	00	00	72	12	52	71	00	03	92	85	95	16	60	72	47	18	00	71	66
501	BLH 1	T	80	50	-	100	10	10	90	40	80	70	10	10	70	90	10	10	30	10	20	80	30	10	10	80
		S	00	00	83	92	17	62	00	60	80	07	68	04	41	56	81	17	44	46	61	13	72	20	52	82
502	BLH 2	T	50	40	70	80	60	80	80	40	80	70	60	60	70	40	60	50	20	60	70	60	70	50	60	80
		S	43	71	78	47	47	40	00	47	70	78	60	25	38	17	59	00	78	74	10	25	34	00	25	71

TABLE 2

The difficulty of 24 test items as rated by teachers and as computed from students' performance on the test items.

CODE	CLASS		TEST ITEMS 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
101	COL 1	T	80	90	30	30	50	40	90	80	70	90	70	60	30	70	90	50	40	30	50	40	90	20	30	90
		S	100	65	70	30	25	10	10	40	30	10	05	20	10	20	35	15	70	25	30	50	30	10	05	15
102	COL 2	T	80	70	30	30	50	40	90	80	70	90	70	60	30	70	90	50	40	30	50	40	90	20	30	90
		S	95	70	70	40	35	15	75	60	75	40	30	20	40	70	60	00	85	40	15	30	60	20	60	15
103	COL 3	T	60	60	30	30	20	20	50	40	20	40	40	40	40	70	70	50	30	30	20	20	20	20	30	50
		S	100	33	43	14	43	14	10	24	57	05	23	10	29	29	10	10	52	38	14	19	14	24	33	19
104	COL 4	T	80	80	60	80	60	10	60	20	50	80	50	10	70	70	20	60	10	10	30	40	20	20	50	70
		S	100	87	79	79	80	04	58	34	83	42	71	25	42	42	25	04	50	75	29	25	17	29	46	38
105	COL 5	T	70	60	60	70	70	30	80	70	70	80	80	30	60	80	80	60	60	50	60	60	60	50	70	50
		S	100	90	65	85	45	30	90	75	70	75	90	35	55	75	70	20	55	50	15	35	20	30	45	35
106	COL 6	T	60	50	40	80	60	20	80	50	80	80	90	20	20	80	20	20	30	30	20	50	20	20	60	60
		S	96	77	46	82	50	23	77	82	86	36	73	05	46	46	72	05	46	68	18	18	18	23	46	27
107	COL 7	T	90	70	70	50	50	60	50	60	60	40	40	80	60	50	40	50	70	40	30	50	60	60	50	80
		S	91	78	82	22	30	34	52	78	87	39	47	73	48	61	39	13	74	57	22	39	17	44	48	65
108	COL 8	T	90	30	10	10	10	10	80	70	70	80	70	60	40	60	10	10	40	20	10	10	10	10	30	70
		S	90	84	63	60	79	39	05	74	79	42	47	32	42	84	84	26	73	79	16	53	32	26	77	63
109	COL 9	T	90	20	100	90	70	100	50	50	50	90	50	80	70	70	70	50	50	30	30	40	70	70	60	80
		S	100	95	65	85	50	30	90	75	75	70	90	35	60	75	75	20	60	55	15	40	25	30	50	40
110	COL 10	T	60	70	60	80	70	70	60	60	70	70	60	50	80	80	70	70	70	60	60	70	70	80	70	80
		S	77	68	86	77	77	27	23	91	96	46	82	18	23	73	55	50	91	77	09	50	46	14	86	46
201	STH 1	T	100	100	80	100	100	100	100	100	100	100	80	50	70	100	100	70		50	70	70	100	80	100	100
		S	90	80	65	50	45	35	30	90	75	15	45	10	45	70	55	75	40	70	25	15	05	30	25	40
202	STH 2	T	70	50	70	80	70	70	90	90	80	90	90	90	80	90	80	80	70	80	70	50	70		70	70
		S	100	70	75	70	40	25	30	85	70	25	45	45	30	60	55	70	60	60	15	25	05	25	55	50
203	STH 3	T	80	70	60	80	70	80	70	60	70	90	90	70	90	70	70	70		60	70	70	90	70	100	70
		S	95	85	75	55	30	30	50	85	70	25	50	45	45	85	60	70	50	60	20	20	05	30	70	50
204	STH 4	T	60	80	80	80	80	80	70	80	80	80	80	80	80	80	80	60	80	80	40	60	80	90	80	50
		S	95	100	85	85	40	90	85	80	95	70	55	65	35	85	60	35	75	80	30	30	60	10	60	65
205	STH 5	T	60		50	50	70	80	50	60	70	40	40	60	70	60	40	60		30	80	60	50	60	70	40
		S	95	100	80	90	40	90	85	75	95	75	65	60	35	80	65	40	75	85	30	30	50	05	55	65
206	STH 6	T	100	100	70	100	80	100	100	100	100	100	80	80	50	100	100	100	100	70	80	60	100	30	100	100
		S	100	75	65	70	70	70	35	55	80	75	80	25	60	75	70	40	80	80	65	40	40	15	55	60
301	MAN 1	T	50	30	10	30	20	30	30	20	40	20	30	10	30	40	100	100	100	70	80	60	100	30	100	100
		S	80	53	53	87	60	87	53	47	80	40	53	13	07	13	60	13	40	80	13	33	67	13	53	40
302	MAN 2	T	80	60	90	80	70	80	80	90	80	80	70	30	40	70	70	60	80	80	60	40	60	20	60	80
		S	94	88	65	88	82	94	82	77	82	71	82	18	24	29	100	29	59	88	18	47	47	18	35	71
401	BUC 1	T	80	70	40	60	60	60	80	60	70	70	60	40	50	50	70	50	40	50	50	30	30	30	20	40
		S	100	90	80	85	45	70	45	75	80	80	90	35	35	95	70	40	95	90	60	40	35	40	75	55
402	BUC 2	T	99	100	60	100	90	100	100	90	100	100	90	100	100	100	100	90	80	90	70	90	100	90	100	100
		S	100	100	90	79	32	100	74	79	95	74	90	79	53	74	95	21	84	95	37	42	79	21	90	58
501	BLH 1	T	70	70	-	80	30	20	60	70	80	70	20	10	90	80	10	10	30	30	20	70	20	20	50	80
		S	100	100	92	62	79	87	00	79	87	58	58	08	75	42	46	12	50	79	42	25	50	39	58	50
502	BLH 2	T	50	40	60	70	70	60	60	50	50	60	80	40	70	60	50	50	30	70	50	60	70	60	60	70
		S	68	40	84	44	48	76	00	44	64	64	36	08	40	80	24	08	72	88	20	44	20	08	40	32

TABLE 3

The discriminating power of each 24 test items as predicted by teachers and as calculated from students' performance of the items

CODE	CLASS		TEST ITEMS 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
101	COL 1	T	40	60	60	40	40	30	60	50	40	60	50	50	50	40	60	50	40	30	20	30	50	20	30	50
		S	00	36	49	43	43	92	00	86	73	23	14	15	54	29	67	00	73	80	10	49	00	..	1.	..
102	COL 2	T	40	60	60	40	40	30	60	50	40	60	50	50	50	40	60	50	40	30	20	30	50	20	30	50
		S	80	68	68	00	58	00	00	43	20	38	00	28	42	89	46	00	00	81	27	00	57	32	33	24
103	COL 3	T	50	60	40	40	40	60	80	80	40	80	60	90	60	60	80	60	70	40	50	80	70	60	60	60
		S	00	53	20	38	62	01	00	21	30	56	00	14	76	62	14	13	98	26	41	50	00	00	53	15
104	COL 4	T	40	60	60	60	60	10	60	30	60	60	50	20	60	60	40	40	10	20	30	30	40	20	40	50
		S	00	20	100	43	21	00	00	17	94	100	44	15	00	24	106	00	22	14	00	30	00	00	20	45
105	COL 5	T	20	50	40	20	40	40	40	50	50	30	20	30	50	80	50	40	50	50	70	30	30	30	20	50
		S	00	29	29	57	03	63	00	51	42	46	00	32	83	29	57	37	06	09	15	51	00	00	49	03
106	COL 6	T	60	70	60	50	50	20	70	40	70	70	60	20	20	60	20	B0	80	50	20	60	20	20	?0	90
		S	32	50	43	00	38	44	00	47	51	68	41	00	20	43	38	00	28	38	32	10	00	47	1.	71
107	COL 7	T	10	40	50	70	70	80	90	90	70	100	90	30	40	60	100	60	40	100	100	60	90	40	50	20
		S	00	38	66	63	07	00	55	51	39	43	05	00	19	100	16	49	200	00	00	20	29	08	35	60
108	COL 8	T	20	40	10	10	10	10	60	60	60	60	70	80	40	30	20	10	40	20	20	20	20	20	30	60
		S	00	24	B9	97	100	43	14	6B	29	72	74	00	00	2B	41	81	6B	15	00	17	43	39	..	14
109	COL 9	T	50	90	80	60	B0	40	60	70	90	90	80	60	70	70	80	90	90	100	100	B0	70	70	80	90
		S	12	29	20	43	29	74	00	40	39	24	00	29	73	32	43	47	17	17	13	61	00	00	42	00
110	COL 10	T	50	50	50	30	50	50	70	50	50	50	70	70	30	50	50	50	30	50	50	50	50	80	50	30
		S	04	33	72	43	10	57	00	31	40	70	43	05	20	73	70	62	14	48	00	04	72	14	31	95
201	STH 1	T	10	20	40	10	10	20	10	10	10	10	30	30	50	10	10	50	-	40	30	40	10	20	10	10
		S	19	66	45	73	24	24	00	19	42	15	14	14	34	41	45	36	62	58	32	00	00	22	53	48
202	STH 2	T	60	70	70	60	50	50	80	60	60	40	40	30	80	50	60	B0	80	70	70	30	60	-	30	60
		S	00	90	04	47	57	42	00	05	95	10	73	32	43	46	48	00	40	57	32	00	00	22	53	48
203	STH 3	T	100	80	90	60	90	90	70	80	100	90	80	90	B0	90	90	40	-	80	100	90	70	70	20	90
		S	33	78	22	71	75	7B	00	17	71	57	65	00	49	12	58	14	44	86	00	00	00	13	52	50
204	STH 4	T	90	90	90	90	90	70	90	90	90	90	90	90	90	70	90	70	90	90	90	70	90	90	90	90
		S	61	90	15	9B	00	55	00	55	27	61	59	34	13	B5	57	50	50	49	53	18	82	00	54	59
205	STH 5	T	60	-	60	B0	50	40	90	B0	40	B0	B0	90	70	B0	B0	70	-	70	100	60	80	40	70	100
		S	64	95	15	14	00	58	00	52	29	71	55	32	13	59	43	46	45	52	45	16	98	00	48	64
206	STH 6	T	10	10	40	10	30	10	10	10	10	10	30	50	30	10	10	10	10	30	50	70	10	10	10	50
		S	00	81	63	30	19	57	00	B1	34	29	00	57	50	32	56	69	24	00	43	01	55	00	76	..
301	MAN 1	T	30	30	10	40	30	40	40	30	30	30	40	10	40	30	10	10	10	30	50	70	10	10	10	50
		S	91	27	27	49	66	B9	00	59	47	100	49	01	00	01	65	00	19	33	31	24	36	18	00	67
302	MAN 2	T	70	70	90	60	80	60	100	60	80	100	60	60	60	80	60	60	60	70	80	70	80	20	60	70
		S	59	58	36	00	98	27	00	57	57	70	36	00	38	76	01	03	35	00	04	51	51	73	35	40
401	BUC 1	T	100	90	40	80	60	80	70	70	90	70	60	20	30	50	50	50	40	60	80	20	20	20	20	50
		S	..	40	01	28	07	77	54	B6	56	57	29	62	10	40	67	65	58	18	58	34	62	17	81	72
402	BUC 2	T	10	10	30	10	30	10	10	30	10	10	30	10	10	10	10	30	30	30	50	30	10	30	10	10
		S	00	00	30	33	00	00	00	50	12	60	33	00	03	83	17	50	16	16	67	54	18	34	33	..
501	BLH 1	T	90	70	-	60	60	50	70	B0	90	70	90	90	60	60	10	10	60	80	60	60	50	50	70	70
		S	00	00	71	29	1B	02	00	75	0B	07	92	04	45	68	43	17	49	52	78	13	72	21	60	72
502	BLH 2	T	60	60	70	90	70	70	B0	70	B0	80	80	100	90	80	100	90	60	80	70	80	90	70	80	80
		S	48	45	15	54	53	44	00	54	99	26	75	26	41	25	73	00	58	57	10	26	36	00	26	00

TABLE 4

The validity of each 24 test items as rated by teachers and as calculated from students' performance on the items

TEST ITEMS

CODE	CLASS		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
101	COL 1	T	90	100	90	90	90	90	90	-	70	90	50	90	00	90	90	90	70	90	30	70	90	30	90	70
		S	00	34	44	73	78	68	00	76	59	22	93	93	48	75	56	20	59	05	10	44	19	19	93	19
102	COL 2	T	100	60	10	20	20	20	100	60	60	90	70	60	10	60	100	40	20	20	50	20	90	20	20	80
		S	63	56	51	17	50	23	00	80	19	35	00	27	38	67	42	00	14	63	26	38	80	31	31	23
103	COL 3	T	70	60	10	10	10	10	80	40	50	60	30	70	20	80	60	30	30	20	30	30	40	20	20	50
		S	33	47	20	36	53	01	00	20	29	49	05	14	61	53	73	13	70	26	38	45	57	19	47	15
104	COL 4	T	100	90	60	60	60	10	60	40	60	100	80	20	60	60	20	40	10	10	30	50	20	20	40	60
		S	00	19	72	40	21	00	00	17	69	71	41	15	02	23	73	00	21	14	12	28	24	11	20	41
105	COL 5	T	80	80	50	60	60	-	60	40	50	00	50	20	50	60	60	20	50	60	50	20	50	30	60	80
		S	00	81	28	92	03	53	00	46	39	42	76	31	64	28	50	34	06	09	15	46	16	15	44	03
106	COL 6	T	80	80	80	80	70	60	70	40	90	80	80	20	70	80	40	80	80	80	60	50	10	60	90	90
		S	30	45	39	87	35	40	00	43	45	56	38	08	20	39	35	00	27	73	31	10	04	43	18	58
107	COL 7	T	90	90	90	90	50	70	50	50	90	80	50	60	60	60	60	70	60	90	60	60	100	60	90	60
		S	104	89	55	53	02	14	48	45	37	84	05	24	19	80	16	44	20	21	07	20	83	08	33	52
108	COL 8	T	100	20	10	10	10	10	90	60	70	90	80	70	20	30	20	10	10	20	10	10	20	20	20	80
		S	87	71	67	70	71	81	25	56	73	91	59	00	38	27	38	63	56	76	12	17	73	36	00	77
109	COL 9	T	100	60	100	100	80	90	80	40	50	60	60	70	60	30	70	60	60	40	30	50	70	40	30	60
		S	12	88	19	93	28	59	00	37	34	23	73	28	59	31	39	44	16	17	13	52	26	28	39	08
110	COL 10	T	50	50	10	50	50	50	70	30	70	70	70	30	50	70	70	70	30	70	50	70	70	10	10	70
		S	04	31	58	72	10	49	00	24	37	58	90	05	20	59	57	53	74	43	02	04	89	78	30	69
201	STH 1	T	10	40	50	20	20	20	10	10	10	10	30	10	40	10	40	50	-	40	30	40	10	10	30	20
		S	19	55	41	59	24	23	00	19	39	81	81	14	32	38	41	29	53	50	31	40	00	22	47	11
202	STH 2	T	60	70	70	70	80	80	90	70	60	70	80	70	80	80	100	90	70	70	70	80	80	-	80	50
		S	00	67	04	42	50	39	00	05	69	86	59	30	40	42	43	00	37	93	31	23	00	21	47	44
203	STH 3	T	100	60	90	60	00	90	60	90	100	90	90	60	90	90	90	60	-	90	100	90	90	90	90	90
		S	31	61	21	58	60	61	00	15	81	79	54	00	44	12	50	14	41	87	00	36	00	13	46	35
204	STH 4	T	90	90	90	90	90	90	90	90	90	90	90	70	90	90	70	50	90	90	50	70	90	90	90	90
		S	52	67	83	70	28	48	00	48	26	89	51	32	13	65	75	44	44	44	47	18	63	48	40	51
205	STH 5	T	50	-	70	90	60	60	90	40	90	50	80	30	80	80	60	80	-	70	100	40	90	40	80	100
		S	54	69	80	72	00	50	00	46	28	83	48	30	13	51	76	42	41	46	41	16	70	00	43	54
206	STH 6	T	50	50	50	50	50	50	50	30	30	30	50	30	50	50	50	30	30	50	50	30	50	10	50	50
		S	43	63	53	29	18	80	00	63	32	76	103	50	45	30	49	57	23	00	77	01	48	00	48	69
301	MAN 1	T	50	-30	50	80	60	80	80	40	80	40	80	20	80	90	50	30	30	50	50	30	50	10	50	50
		S	67	27	26	44	81	66	00	51	43	97	71	100	13	01	56	22	18	83	30	24	34	17	06	82
302	MAN 2	T	100	90	80	80	80	60	90	60	80	80	100	30	60	90	70	40	60	80	100	10	70	20	70	80
		S	51	51	34	00	70	26	00	50	49	57	34	00	35	61	01	03	65	108	04	45	45	59	33	37
401	BUC 1	T	60	80	20	80	40	90	80	60	80	50	60	20	50	50	80	50	20	70	90	30	30	20	20	70
		S	00	37	01	77	06	61	48	96	49	84	83	53	10	37	56	54	50	18	56	32	53	16	63	82
402	BUC 2	T	90	90	90	90	90	90	70	30	90	90	70	50	90	90	60	50	10	30	70	30	90	30	70	90
		S	00	00	29	32	03	00	00	72	12	52	71	00	03	92	85	95	16	60	72	47	18	32	71	66
501	BLH 1	T	80	40	-	80	20	10	70	40	90	50	10	10	70	80	10	10	20	50	10	90	10	10	10	90
		S	00	00	83	92	17	02	00	60	80	07	68	04	41	56	81	17	44	46	61	13	72	20	52	82
502	BLH 2	T	40	60	60	80	60	60	60	30	60	50	40	40	60	40	40	50	20	60	50	50	50	40	60	70
		S	43	71	78	47	47	40	00	47	70	70	60	25	38	17	59	00	78	74	10	25	34	00	25	71

TABLE 5

The test -retest stability of 24 test items as as predicted by teachers and as actually calculated from students' performance on the items

TEST ITEMS

CODE	CLASS		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
101	COL 1	T	80	80	100	90	80	80	00	90	80	100	80	90	50	70	90	60	80	90	80	80	100	70	80	90
		S	75	62	71	33	20	50	50	55	66	50	100	75	50	50	71	67	79	100	50	60	17	100	100	33
102	COL 2	T	80	80	100	90	80	80	100	90	80	100	80	80	50	70	90	60	80	90	80	80	100	70	80	90
		S	100	92	93	88	67	67	93	89	82	100	62	100	100	100	64		94	86	00	83	100	50	92	80
103	COL 3	T	80	60	100	100	100	80	80	70	50	70	80	80	40	100	60	40	70	80	80	80	100	70	80	70
		S	95	57	67	67	56	100	00	67	54	100	75	50	67	43	50	100	55	78	100	50	100	60	63	00
104	COL 4	T	90	80	90	100	90	90	90	30	50	90	80	70	80	70	60	70	60	60	60	70	60	60	60	70
		S	100	95	90	90	88	100	85	77	95	100	82	60	53	78	100	100	92	89	43	33	25	57	100	11
105	COL 5	T	70	60	60	90	90	100	80	90	80	80	80	80	60	80	70	70	80	80	60	90	70	80	80	80
		S	100	88	85	85	67	60	100	85	92	87	94	63	56	93	92	75	80	100	100	50	60	33	57	93
106	COL 6	T	90	70	70	80	50	30	70	50	70	90	90	20	20	70	20	60	70	50	20	60	30	20	60	80
		S	100	100	100	100	63	33	100	100	79	71	85	100	25	67	92	100	85	80	67	100	100	33	57	67
107	COL 7	T	100	100	100	90	80	100	100	100	90	100	100	100	100	90	100	100	90	100	100	80	90	100	90	100
		S	100	94	89	100	67	71	100	100	95	75	90	94	82	79	89	100	88	86	40	100	100	60	100	67
108	COL 8	T	100	80	90	90	90	90	90	70	80	90	80	70	60	60	80	30	70	70	50	20	50	50	50	90
		S	80	100	100	100	81	88	100	93	87	88	100	100	88	87	94	83	100	93	100	91	83	40	94	25
109	COL 9	T	90	50	90	100	100	100	100	80	90	100	100	100	90	90	90	80	80	70	80	90	90	80	80	90
		S	100	88	100	79	67	60	100	85	92	87	94	64	56	93	93	75	80	100	100	50	60	33	57	63
110	COL 10	T	100	88	100	79	67	60	100	85	92	87	94	63	56	93	93	75	80	100	100	50	60	33	51	83
		S	88	73	100	100	100	67	67	100	100	90	94	67	83	94	92	80	100	94	00	92	91	00	94	90
201	STH 1	T	100	100	100	100	100	100	100	100	100	100	100	80	100	100	100	100		100	80	90	100	100	100	100
		S	94	100	100	100	88	67	83	94	80	33	78	100	78	100	82	87	88	93	00	100	00	40	93	75
202	STH 2	T	100	90	100	100	90	80	90	100	90	100	100	90	100	80	90	100	80	90	80	60	80		10	20
		S	100	100	100	100	75	50	86	100	93	80	70	100	43	100	70	86	73	83	50	40	100	17	100	90
203	STH 3	T	80	40	90	90	90	90	90	90	90	100	100	90	100	100	90	100		90	90	100	100	90	100	90
		S	100	100	100	100	86	60	89	100		93	50	100	100	63	100	80	93	73	83	33	50	00	17	100
204	STH 4	T	90	90	90	80	80	80	80	90	80	80	80	70	90	80	80	80	80	80	80	90	80	90	100	80
		S	100	94	93	87	88	100	100	86	88	85	91	00	63	73	82	86	85	87	83	40	100	50	83	91
205	STH 5	T	100		90	90	90	100	70	100	100	80	80	80	100	70	60	80		70	80	90	90	100	100	70
		S	94	94	93	88	88	88	80	71	77	57	82	52	43	73	73	75	97	86	83	43	100	100	77	83
206	STH 6	T	100	100	70	100	80	100	100	100	100	100	80	80	50	100	100	100	100	70	80	60	100	30	100	100
		S	100	86	92	100	86	100	88	91	88	100	100	60	73	100	100	100	94	100	86	100	86	67	100	92
301	MAN 1	T	100	100	100	100	100	100	100	100	100	90	100	50	100	100	100	100	100	100	100	100	100	50	100	100
		S	100	100	75	100	75	100	100	67	82	100	86	50	00	00	86	100	67	90	00	75	90	50	100	67
302	MAN 2	T	100	90	90	90	90	80	90	100	100	100	100	30	50	90	80	80	80	100	90	80	100	30	80	100
		S	100	95	90	100	85	100	100	92	92	91	92	50	100	100	86	100	89	100	67	83	100	33	60	23
401	BUC 1	T	90	90	60	90	90	80	80	70	80	70	70	50	50	60	70	60	50	60	70	30	30	30	20	50
		S	100	100	94	100	78	93	88	80	94	100	94	71	86	95	86	100	95	94	100	88	100	88	100	91
402	BUC 2	T	100	100	100	100	90	100	100	90	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		S	100	100	100	100	100	84	100	87	100	93	88	100	80	93	100	100	94	100	100	75	93	50	94	92
501	BLH 1	T																								
		S																								
502	BLH 2	T																								
		S																								

TABLE 6

The test-retest stability of 24 test items as teachers predicted percentages of pupils in a class who would pass a test item in second testing, on which they have failed on the first occasion

TEST ITEMS

CODE	CLASS		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
101	COL 1	T	90	40	20	40	30	40	90	80	60	80	50	40	20	40	80	40	50	20	10	30	80	10	30	40
		S	00	14	50	21	33	39	17	33	29	17	05	19	06	33	15	06	83	53	36	30	07	00	00	24
102	COL 2	T	80	40	20	40	30	40	90	80	60	80	50	40	20	40	80	40	50	20	10	30	80	10	30	40
		S	00	33	20	08	08	31	40	13	75	17	23	00	36	17	13		100	25	06	15	25	20	00	25
103	COL 3	T	10	10	10	10	10	10	20	10	30	20	20	20	10	20	30	10	10	10	10	10	10	20	20	10
		S	00	29	08	11	08	00	16	13	13	05	12	00	00	07	07	00	30	67	11	12	05	18	08	17
104	COL 4	T	30	30	30	40	30	10	50	20	20	40	20	30	20	20	20	20	30	30	20	30	30	30	40	30
		S	00	67	50	25	08	22	18	09	80	13	71	26	08	20	11	00	25	00	06	11	10	35	23	33
105	COL 5	T	30	10	20	20	20	10	10	20	20	20	20	20	10	10	20	20	20	20	30	20	20	20	20	20
		S	00	00	29	20	11	23	00	40	80	100	100	10	22	50	80	07	50	00	06	17	00	33	20	58
106	COL 6	T	10	10	20	10	20	20	20	20	10	10	20	10	10	20	10	20	20	20	20	10	10	10	10	20
		S	10	50	82	50	11	36	00	33	00	40	75	75	00	00	25	81	71	80	50	23	71	50	60	43
107	COL 7	T	90	50	50	50	50	60	50	80	50	70	40	90	70	50	40	40	70	50	40	50	40	50	30	60
		S	50	17	60	39	12	25	08	83	100	27	23	00	17	22	14	00	43	56	22	20	15	39	50	38
108	COL 8	T	10	80	10	10	10	10	20	20	10	30	20	20	20	20	10	20	10	10	10	20	20	20	20	30
		S	50	00	33	14	00	36	05	00	00	00	10	08	09	50	00	15	00	50	00	25	30	21	100	29
109	COL 9	T	100	50	100	50	50		50	50	50	50	90	60	80	50	50	40	60	30	30	40	60	30	30	60
		S	00	00	38	25	11	23	00	40	80	100	100	10	21	50	80	07	50	00	06	16	00	33	09	58
110	COL 10	T	20	20	20	20	10	20	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
		S	20	14	00	25	25	25	19	50	00	08	00	05	25	40	11	00	00	60	05	20	00	00	33	16
201	STH 1	T	10	10	50	10	10	10	10	10	10	10	50	20	20	10	10	30		20	20	20	10	10	20	10
		S	50	10	71	36	25	29	29	00	20	11	09	42	08	50	67	20	42	50	00	06	06	07	60	50
202	STH 2	T	30	20	10	20	10	20	10	20	10	10	10	20	10	20	10	30	30	20	20	30	2L		10	20
		S	00	40	50	17	33	29	39	33	17	20	10	30	23	25	50	33	44	25	12	27	00	00	44	50
203	STH 3	T	10	30	10	30	30	20	10	10	10	20	10	10	10	10	10	10		10	10	10	10	10	10	10
		S	00	75	67	33	39	33	27	33	17	13	00	36	17	25	60	33	44	38	06	19	00	07	50	40
204	STH 4	T	10	10	10	10	10	10	10	10	10	10	10	20	20	10	10	10	10	20	20	20	20	10	20	20
		S	00	00	100	100	20	00	00	50	100	20	71	20	30	30	43	36	60	00	08	31	29	13	50	29
205	STH 5	T	20		30	20	40	20	30	10	10	20	10	10	30	20	10	20		10	20	30	10	30	30	10
		S	00	00	100	100	20	00	00	50	100	25	86	50	46	33	57	40	75	25	08	27	37	12	23	33
206	STH 6	T	10	10	50	20	50	10	50	50	40	40	50	50	50	40	40	50	40	40	50	40	40	50	40	40
		S	00	100	63	00	33	50	25	33	67	00	00	20	22	00	29	09	25	50	33	08	08	06	11	25
301	MAN 1	T	10	10	10	10	10	10	10	10	10	10	10	10	10	10	40	50	40	40	50	40	40	50	40	40
		S	67	57	17	100	50	50	00	00	67	25	29	17	08	08	33	00	25	33	08	10	50	08	17	13
302	MAN 2	T	20	10	10	40	20	20	20	30	20	10	20	10	20	30	20	10	10	20	20	20	30	30	20	20
		S	100	50	00	100	00	00	33	50	33	25	00	08	17	20	00	00	17	00	08	11	29	25	10	25
401	BUC 1	T	50	50	50	50	40	40	40	30	40	40	50	30	50	30	30	20	30	30	30	20	30	30	20	20
		S	00	50	50	100	46	17	00	20	25	50	00	23	15	00	33	08	100	00	13	17	08	00	00	23
402	BUC 2	T	00	10	10					20			10					10	20	10	20	30		10	10	10
		S	00	00	50	75	38	00	20	00	100	60	100	00	00	60	00	13	00	100	33	00	00	07	50	50
501	BLH 1	T																								
		S																								
502	BLH 2	T																								
		S																								

TABLE 7

ITEM PROPERTIES

ITEMS	A	B	C	D	E	F
1	.25 NS	.25 NS	.21 NS	.16 NS	-.045 NS	-.14 NS
2	-.45 S	-.45 S	-.013 NS	-.065 NS	-.033 NS	-.18 NS
3	.21 NS	.21 NS	-.043 NS	-.014 NS	.18 NS	-.11 NS
4	.018 NS	.018 NS	-.16 NS	.32 NS	.014 NS	-.22 NS
5	-.042 NS	-.042 NS	-.11 NS	-.044 NS	-.10 NS	.089 NS
6	.021 NS	.021 NS	.037 NS	.21 NS	-.09 NS	.18 NS
7	.16 NS	.16 NS	.065 NS	.30 NS	.16 NS	.041 NS
8	.13 NS	.13 NS	.31 NS	.19 NS	-.05 NS	.03 NS
9	-.28 NS	-.28 NS	.17 NS	-.22 NS	-.082 NS	.21 NS
10	.008 NS	.008 NS	.40 S	.096 NS	-.06 NS	.25 NS
11	-.007 NS	-.007 NS	-.012 NS	.46 S	.15 NS	-.08 NS
12	-.021 NS	-.021 NS	-.04 NS	.056 NS	.018 NS	-.07 NS
13	.114 NS	.14 NS	.26 NS	-.33 NS	-.085 NS	-.40 S
14	-.031 NS	-.031 NS	.40 S	-.32 NS	-.032 NS	-.05 NS
15	-.32 NS	-.32 NS	.05 NS	-.172 NS	-.038 NS	.36 NS
16	-.031 NS	-.031 NS	.069 NS	.18 NS	-.13 NS	.35 NS
17	-.32 NS	-.32 NS	.019 NS	-.17 NS	.007 NS	.21 NS
18	-.13 NS	-.13 NS	-.093 NS	.08 NS	-.071 NS	.31 NS
19	-.08 NS	-.08 NS	.082 NS	.16 NS	.12 NS	-.18 NS
20	-.23 NS	-.23 NS	-.27 NS	-.36 NS	.16 NS	.18 NS
21	-.08 NS	-.08 NS	.091 NS	.12 NS	-.021 NS	-.085 NS
22	-.18 NS	-.18 NS	-.11 NS	-.28 NS	.011 NS	.20 NS
23	-.08 NS	.08 NS	-.26 NS	.21 NS	-.078 NS	.25 NS
24	-.17 NS	-.17 NS	.24 NS	-.35 NS	.044 NS	.20 NS

KEY: A = Overall quality
B = Difficulty level
C = Discriminating Power
D = Validity
E = Stability 1
F = Stability 2

are positive and the other half are negative. Only 5 of the 144 correlation coefficients are significant, 2 are positive and 3 are negative.

There is no recognizable pattern in these correlation coefficient to indicate whether teachers were better judges of particular item properties, or particular test items. Not only are the correlation coefficients insignificant, but most of them are also near zero. This pattern of the correlation coefficients is not very different from what one would expect of correlation coefficients between two sets of scores randomly selected.

The evidence obtained suggests that teachers, as a group, could not judge test item properties better than chance. The results also indicate that teachers, as a group, were poor judges of all the item properties as well as all the test items.

On the other hand, these results do not give adequate information as to whether some teachers were better judges of test items through student performance than others. The degree of relationship between a single teacher's ratings of the item properties and the values obtained for the item properties from the performance of that particular teacher's class on the 24-item test, is examined in a separate analysis.

2. THE DEGREE OF RELATIONSHIP BETWEEN INDIVIDUAL TEACHERS' RATINGS OF ITEM PROPERTIES AND INDIVIDUAL CLASS PERFORMANCE

On the other hand, to obtain the relationship between ratings of individual teachers and the performance of their respective classes, the two sets of values obtained for each of the 22 classes are correlated. The results of the correlations are shown in Tables 8 and 9.

TABLE 8

Individual teachers' ratings of 6 Psychometric Properties of the items correlated with values obtained for the same properties from performance of their individual classes on the same test items

CLASSES

Psycho-metric INDICES	COL 1	COL 2	COL 3	COL 4	COL 5	COL 6	COL 7	COL 8	COL 9	COL 10	STH 1
A	.120 NS	.300 NS	.41 S	.381 S	.250 NS	-.013 NS	.33 NS	.271 NS	.187 NS	.250 NS	.045 NS
B	.050 NS	.30 NS	.050 NS	.456 S	.630 S	.552 S	.650 S	.112 NS	.003 NS	-.032 NS	.145 NS
C	.120 NS	.110 NS	-.066 NS	.410 S	.056 NS	.184 NS	-.074 NS	-.231 NS	-.225 NS	-.163 NS	.042 NS
D	-.12 NS	.085 NS	.030 NS	.240 NS	-.048 NS	.267 NS	.560 S	.302 NS	.121 NS	.048 NS	.333 NS
E	-.030 NS	.330 NS	.014 NS	.240 NS	-.18 NS	.386 S	-.148 NS	.098 NS	-.070 NS	.245 NS	-.016 NS
F	-.20 NS	.110 NS	.140 NS	.040 NS	-.20 NS	.160 NS	.140 NS	-.141 NS	.180 NS	.003 NS	-.012 NS

TABLE 9

Correlation Coefficients between ratings of individual teachers and the performance of their respective classes over 6 item properties

CLASSES

Psycho-metric INDICES	STH 2	STH 3	STH 4	STH 5	STH 6	MAN 1	MAN 2	BUC 1	BUC 2	BLH 1	BLH 2
A	.161 NS	-.110 NS	.056 NS	.002 NS	.133 NS	.23 NS	-.130 NS	.393 S	.632 S	.0074 NS	-.019 NS
B	.212 NS	-.140 NS	.232 NS	-.023 NS	.411 S	-.025 NS	.715 S	.374 **N S**	.230 NS	.220 NS	.033 NS
C	.072 NS	-.094 NS	-.081 NS	.314 NS	-.028 NS	.220 NS	.005 NS	-.0266 NS	.333 NS	.0162 NS	.005 NS
D	-.210 NS	-.032 NS	.095 NS	-.011 NS	.26 NS	-.110 NS	.049 NS	.290 NS	-.241 NS	.1274 NS	.088 NS
E	.323 NS	-.261 NS	-.290 NS	-.212 NS	.525 S	.200 NS	.643 S	.051 NS	-.034 NS		
F	.072 NS	.248 NS	-.220 NS	-.232 NS	-.23 NS	.342 NS	.470 S	.257 NS	-.050 NS		

3. THE RESULTS

There are 128 correlation coefficients in Tables 8 and 9. Only 14 of the correlation coefficients are significant. Six of the significant correlation coefficients are obtained for item difficulty, 4 for the overall item quality, 2 for item stability and 1 for item discrimination, 1 for item content relevance.

4. CONCLUSION

The results show that teachers were not able to judge item properties adequately through student performance. In six cases out of the 22, judgments were significantly correlated with students' performances. This indicates that teachers judged item difficulty a little better than other item properties. The results are far from encouraging however.

At this stage we can attempt an answer to our second research question. The answer to the question is that teachers cannot judge the properties adequately. This throws some light on the current discussion on teachers' judgments.

This finding was confirmed by a recent study on the accuracy of teachers' forecasting of students' performances. Seddon (1982) found that teachers significantly overestimated the performance of their students. Some of the previous literature reviewed (Mehrens and Lehmann, 1978; Ausubel and Rubinson, 1969; Anastasi, 1969; Lorge and Diamond, 1953; Lucow, Early & Garland, 1976) also pointed to teachers' inability to judge test item properties through perceived student performance. Why are teachers not able to judge item properties adequately? This question has not been investigated in this study.

5. THE RELATIVE CONTRIBUTIONS OF OTHER TEST ITEM PROPERTIES TO THE OVERALL QUALITY OF THE TEST ITEM

Most psychometricians agree that the test item property that weighs most in the assessment of the overall item quality, is the validity of the test item (Flanagan, 1939; Nunally, 1970; Guildford & Grutchter, 1978; Jensen, 1980). The precedence they have given to the item validity is based on evidence obtained from psychometric analysis of student performance on test items. The psychometricians argue that when the validity of the test item is adequate, other test item properties, such as difficulty, discrimination, reliability etc. are also satisfactory (see section on criterion for stating overall item quality).

A subsequent question to be asked is: which item property carries more weight for the assessment of the overall item quality in teachers' judgment of test items? Is it the apparent difficulty of the item in relation to particular pupils, or the items discriminating power between able and less able students, or is it the relevance of the item to what has been taught, etc.?

The item properties in question are the overall quality, the difficulty level, the discriminating power and the stability of the test item. To determine the relative contributions of other item properties to the overall rating of the item, one needs to identify an appropriate procedure.

A multiple regression technique is proposed to obtain the item property that best predicts the criterion variable. One purpose of multiple regression analysis is to discover which independent variable is more related to or predicts or explains more variation in the

dependent variable. A second purpose when the step-wise procedure is used is to rate the variables in order of importance (Snedecor and Cochran, 1967). Another purpose of multiple regression analysis is to disentangle or set aside the effect of a variable(s) and measure the effects of different independent variables on a criterion variable (Mosteller and Tukey, 1977). In other words, the procedure allows the control of other confounding variables in order to evaluate the relative contribution of a variable in question.

Multiple regression analysis is proposed here to find out which item property is more important in determining the overall quality of the item, while the effects of other item properties on the overall item quality is removed. In the present analysis, the overall rating of the item stands as the criterion or dependent variable and the ratings of the remaining five item properties as predictor or independent variables.

6. THE ANALYSIS AND THE RESULTS OF REGRESSION ANALYSIS

6.1. The Analysis

Multiple regression analysis has been used here to summarise and describe the degree of dependence of teachers' ratings of the overall items on their ratings of difficulty (Diff), discrimination (Dis), content relevance (Con Rel), stability 1 (Rel) and stability 2 (Rel 2) of the item. The following symbols are used throughout this chapter:

Qual - Teachers' ratings of the overall item quality

Diff - Teachers' ratings of the difficulty level of the item

Dis - Teachers' ratings of the item's discriminating power

Con Rel - Teachers' ratings of the content relevance of the item

Rel - Teachers' ratings of the reliability of the items, predicting the proportion of candidates who passed the item in the first

testing and would pass the item again in a second testing.

Rel 2 - Teachers' ratings of the reliability of the item, predicting the proportion of candidates who failed the item in the first testing but would pass the item in a second testing.

The purpose of the analysis was to discover which item property is more related to the overall rating (Qual) of the item. Here we are interested in examining the impact of each variable at a time when the effects of other variables on the dependent variable are controlled and when the effects of other independent variables are not controlled. We are concerned with the relationship between each particular independent variable and the dependent (qual), not with the overall dependence of Qual on Diff, Dis, Con Rel, Rel and Rel 2.

First the overall rating, Qual, was regressed on the other five item properties. The order in which the predictor variables entered into the equation was predetermined. The researcher has specified the order of inclusion. The order of inclusion of variables in the equation is as follows: Diff, Dis, Con Rel, Rel and Rel 2. However, this order of inclusion of variables does not necessarily imply any prior knowledge of the relative contributions by these variables to the explained variance of the dependence variable, Qual.

7. THE RESULTS

The results of the regression analysis are shown in Tables 10 to 22. Tables 10 to 15 depict the results of the analysis when the order of inclusion of the independent variables in the procedure was predetermined. On the other hand, Tables 17 to 22 concern the results of regression analysis when the order of inclusion was determined by the respective contributions of each variable to the explained variance.

TABLE 10

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	COL 1 REGRESSION B	MULTIPLE R	R^2	F
DIFF	.0452	.842		-.1494	.3189	.1017	.138
DIS	.0958	.672		.3185	.3280	.1076	.127
CON REL	.2759	.214		.2561	.4024	.1619	.057
REL	.0779	.730		-.7732	.1300	.0169	.554
REL 2			REL2	.2366	.4713	.2221	.023

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	COL 2 REGRESSION B	MULTIPLE R	R^2	F
DIFF	-.3467	.105		-.2848	.5001	.2501	.013
			DIS	.6175	.5213	.2717	.009
CON REL	.0583	.792		.1435	.3077	.0947	.144
REL	.3517	.100		.5507	.4575	.2093	.025
REL 2	.2929	.175		.2924	.4956	.2456	.014

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	COL 3 REGRESSION B	MULTIPLE R	R^2	F
DIFF	.1136	.606		.1735	.2792	.0780	.183
DIS	.1415	.520		.2708	.3194	.1020	.128
CON REL			CON REL	.2301	.4108	.1687	.046
REL	.4622	.026		.2745	.3729	.1391	.073
REL 2	-.1652	.451		.9205	.0472	.0022	.827

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	COL 4 REGRESSION B	MULTIPLE R	R^2	F
DIFF	-.3454	.106		.1512	.0279	.0010	.897
DIS	.1174	.594		.4687	.5721	.3273	.003
CON REL			CON REL	.3529	.6780	.4600	.000
REL	.3932	.063		.5131	.5894	.3474	.002
REL 2	-.2195	.314		.4972	.0345	.0012	.873

Multiple R,R^2 and regression coefficients before the effect of any variable is removed. The order of inclusion of variables in the equation is that in which the variables appeared on the scale

TABLE 11

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	COL 5 REGRESSION B	MULTIPLE R	R^2	F
DIFF	-.2977	.168		.1908	.3491	.1218	.095
DIS	.0919	.677		.7210	.0871	.0076	.686
CON REL			CON REL	.2644	.4763	.2269	.019
REL	.2147	.325		.3339	.0033	.0000	.988
REL2	.4357	.058		-.76164	.3249	.1055	.121

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	COL 6 REGRESSION B	MULTIPLE R	R^2	F
DIFF	-.2080	.341		-.9000	.1942	.0223	.486
DIS	.1360	.536		.3220	.5217	.2722	.009
CON REL	.		CON REL	.5585	.6851	.4693	.000
REL	-.0633	.774		.2358	.3975	.1580	.054
REL2	.1299	.554		.5000	.1759	.0309	.411

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	COL 7 REGRESSION B	MULTIPLE R	R^2	F
DIFF	-.0028	.990		-.4878	.0641	.0041	.766
DIS	-.0640	.772		-.5688	.0932	.0087	.665
CON REL			CON REL	.4395	.4381	.1920	.032
REL	.0709	.748		.1356	.0534	.0029	.804
REL2	.0136	.951		-.5172	.0495	.0025	.818

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	COL 8 REGRESSION 8	MULTIPLE R	R^2	F
DIFF	.6309	.001		.6054	.4607	.2122	.023
DIS	.5762	.004		1.1773	.8294	.6880	.000
CON REL			CON REL	.9199	.9474	.8975	.000
REL	-.0086	.969		.6745	.4489	.2016	.028
REL2	.2115	.333		.3275	.1509	.0223	.482

Multiple R, R^2 and Regression Coefficients before the effect of any variable is removed. The order of inclusion of variables in the equation is that in which the variables appeared on the scale.

TABLE 12

COL 9

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	REGRESSION B	MULTIPLE R	R^2	F
OIFF	-.2460	.258		-.3587	.5482	.3005	.006
OIS	-.0811	.713		-.6250	.4548	.2063	.026
CON REL			CON REL	.7268	.7515	.5648	.000
REL	-.1536	.484		.4550	.2560	.0655	.227
REL2	.0719	.744		.3216	.3513	.1234	.092

COL 10

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	REGRESSION B	MULTIPLE R	R^2	F
OIFF	.0218	.921		-.6522	.0903	.0082	.675
OIS	.1490	.487		.2553	.2527	.0638	.234
CON REL			CON REL	.3484	.6261	.3919	.001
REL	.0286	.897		.1483	.2387	.0570	.261
REL2	-.1544	.482		-.7579	.2666	.0711	.208

STH 1

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	REGRESSION B	MULTIPLE R	R^2	F
OIFF	-.447	.088		.2976	.6069	.3683	.002
OIS	.6507	.001		1.0273	.8751	.7658	.000
CON REL			CON REL	.9876	.8511	.7243	.000
REL	.2394	.283		-.3846	.1380	.0910	.530
REL2	.1348	.550		.8662	.6138	.3770	.002

STH 2

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	REGRESSION B	MULTIPLE R	R^2	F
OIFF	.0436	.847		-.2183	.3885	.1510	.057
OIS	.1401	.534		.2807	.0348	.0012	.875
CON REL			CON REL	.2264	.1882	.0354	.390
REL	-.2028	.365		-.1771	.0325	.0011	.883
REL2	-.1500	.505		-.4818	.2727	.0743	.208

Multiple R,R^2 and Regression Coefficients before the effect of any variable is removed. The order of inclusion of variables in the equation is that in which the variables appeared on the scale

TABLE 13

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	STH 3 REGRESSION B	MULTIPLE R	R^2	F
DIFF	.2220	.321		-.1182	.0219	.0005	.921
DIS	.4174	.053		.1758	.2558	.0654	.239
CON REL	.1391	.537		.1235	.2367	.0560	.277
REL			REL	.5784	.6151	.3873	.002
REL 2	-.2500	.262		-.7791	.4771	.2276	.021

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	STH 4 REGRESSION B	MULTIPLE R	R^2	F
DIFF			DIFF	-.5435	.8718	.7601	.000
DIS	.1395	.525		.3500	.2618	.0685	.217
CON REL	.5080	.013		.6186	.7585	.5753	.000
REL	-.0417	.850		-.7143	.0447	.0020	.886
REL 2	-.2229	.307		-.5000	.2365	.0559	.266

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	STH 5 REGRESSION B	MULTIPLE R	R^2	F
DIFF	.0992	.669		.1404	.1108	.0123	.624
DIS	.3890	.081		.5625	.4970	.2471	.019
CON REL			CON REL	.5506	.5749	.3305	.005
REL	-.1730	.453		-.3940	.2468	.0609	.268
REL 2	-.2758	.226		-.6111	.2808	.0788	.206

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	STH 6 REGRESSION B	MULTIPLE R	R^2	F
DIFF	-.415	.049		-.2500	.3664	.1343	.078
DIS	-.09140	.678		-.4451	.0375	.0014	.862
CON REL			CON REL	1.3197	.7148	.5109	.000
REL	.3634	.088		.5289	.4825	.2328	.000
REL 2	-.2152	.324		-.5748	.3608	.1302	.083

Multiple R,R^2 and regression coefficients before the effect of any variable is removed. The order of inclusion of variation in the equation is that in which the variables appeared on the scale

TABLE 14

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	MAN 1 REGRESSION B	MULTIPLE R	R^2	F
DIFF	-.0278	.900		.1876	.2871	.0824	.174
DIS	.0863	.695		.4206	.3390	.1149	.105
CON REL			CON REL	.7378	.8100	.6561	.000
REL	.4094	.052		.91025	.6387	.4080	.0001
REL2	.0292	.895		-.4800	.4025	.1620	.051

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	MAN 2 REGRESSION B	MULTIPLE R	R^2	F
DIFF	.1829	.404		.4174	.0365	.0013	.865
DIS	.4595	.029		.8704	.6392	.4086	.001
CON REL			CON REL	.6274	.6609	.4367	.000
REL	.1092	.620		.5844	.5171	.2674	.010
REL 2	.0488	.825		.1093	.0360	.0013	.868

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	BUC 1 REGRESSION B	MULTIPLE R	R^2	F
DIFF	.1926	.379		.1953	.2072	.0430	.331
DIS	.2515	.247		.45B3	.6316	.3990	.001
CON REL			CON REL	.4919	.6481	.4200	.001
REL	-.1783	.416		.3168	.0138	.0002	.949
REL 2	.1278	.561		.4655	.2723	.0742	.198

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	BUC 2 REGRESSION B	MULTIPLE R	R^2	F
DIFF	-.5444	.044		-.2486	.3324	.1105	.226
DIS	.2009	.491		.7188	.3866	.1495	.155
CON REL			CON REL	-.1322	.0162	.0003	.954
REL	.2726	.364		-2.6429	.2973	.0884	.282
REL 2	.0651	.825		.8151	.2961	.0877	.284

Multiple R, R^2 and regression coefficients before the effect of any variable is removed. The order of inclusion of variables in the equation is that in which the variables appeared on the scale

TABLE 15

Multiple R,R^2 and regression coefficients before the effect of any variable is removed. The order of inclusion of variables in the equation is that in which the variables appeared on the scale

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR. REMOVE	REGRESSION B	MULTIPLE R	R^2	F
DIFF	.0262	.908		.2065	.1203	.0145	.585
DIS	-.1438	.523		.4457	.2790	.0778	.197
CON REL			CON REL	.9624	.9198	.8461	.000
REL							
REL 2							

VARIABLE INCLUSION	PARTIAL CORR.	F	VAR.REMOVE	REGRESSION B	MULTIPLE R	R^2	F
DIFF	.1119	.611		-.8271	.0903	.0082	.675
DIS	.4042	.056		.5568	.4036	.1630	.050
CON REL			CON REL	.8800	.7506	.5635	.000
REL							
REL 2							

The regression was done separately for each teacher's scores against his/her 30+ pupils.

The purpose of the analysis of teachers' judgments of the test item properties was to discover the relationships between teachers' overall ratings of the test items and each of their ratings of the remaining five test item properties. When the order of the inclusion of the independent variables is arbitrarily imposed (see Tables 10 to 15), content relevance of the item to what has been taught is significantly related to the overall rating for 18 teachers out of 22; item difficulty has significant relationship with the overall rating for 6 teachers out of 22; item discrimination has significant relationship with the overall rating for 10 teachers out of 22. Stability 1 and stability 2 have significant relationships with the overall rating for 8 teachers out of 20 and 5 teachers out of 20 respectively. The results are summarized in Table 16.

TABLE 16

Variable	No. of teachers	No. of sign relations	Percent
Diff	22	6	27
Dis	22	10	45
Con Rel	22	18	82
Rel	20	8	40
Rel 2	20	5	25

Tables 10 to 15 also present partial coefficients, and their statistical significance, between the dependent variable (Qual) and 4 independent variables when component of the variance due to the most important independent variable is removed. Content relevance (Con Rel)

TABLE 17

The Summary Table of the Stepwise Multiple Regression Analysis for 1' teacher. Order of inclusion of variables is determined by the respective contribution of each variable to explained variance

COL 1

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	REL 2	.4513	.2221	.2221	.023
2	CON REL	.5304	.2813	.0592	.037
3	REL	.5362	.2875	.0062	.086
4	DIFF	.5416	.2933	.0058	.160
5	DIS	.5444	.2964	.0030	.263

COL 2

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	DIS	.5213	.2717	.2717	.009
2	REL	.5015	.3618	.0901	.009
3	DIFF	.6281	.3945	.0323	.016
4	CON REL	.6451	.4162	.0217	.030
5	REL 2	.6859	.4704	.0543	.031

COL 3

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.4108	.1674	.1687	.046
2	REL 2	.5885	.3463	.1776	.012
3	DIFF	.6504	.4231	.0767	.010
4	DIS	.6761	.4571	.0341	.16
5	REL 2	.6770	.4584	.0012	.036

COL 4

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.6780	.4600	.4600	.000
2	REL	.7370	.5432	.0836	.000
3	REL 2	.7818	.6112	.0680	.000
4	DIS	.8013	.6420	.0309	.000
5	DIS	.8300	.6900	.0469	.000

TABLE 18

The summary table of the stepwise multiple regression analysis for 1 teacher. Order of inclusion of variable is determined by the respective contribution of each variable to explained variances

COL 5

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.4763	.2669	.2669	.019
2	REL 2	.6113	.3736	.1467	.007
3	DIFF	.6392	.4086	.0350	.013
4	DIS	.6570	.4317	.0231	.024
5	REL	.6687	.4471	.0155	.043

COL 6

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.6851	.4693	.4693	.000
2	DIFF	.7016	.4923	.0230	.000
3	DIS	.7266	.5280	.0357	.002
4	REL	.7514	.5647	.0367	.002
5	REL 2	.7525	.5663	.0016	.006

COL 7

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.4381	.1920	.1920	.032
2	REL 2	.4428	.1960	.0041	.101
3	REL 2	.4466	.1994	.0034	.207
4	DIFF	.4543	.2064	.0070	.329
5					

COL 8

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.9474	.8975	.8975	.000
2	DIFF	.9687	.9383	.0408	.000
3	DIS	.9696	.9420	.0019	.000
4	REL	.9702	.9412	.0011	.00
5	REL 2				

TABLE 19

The summary table of the stepwise multiple regression analysis for one teacher. Order of inclusion of variable is determined by the respective contribution of each variable to explained variances.

COL 9

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.7515	.5648	.5647	.000
2	DIFF	.7688	.5911	.0263	.000
3	REL	.7733	.5980	.0069	.000
4	REL 2	.7829	.6129	.0090	.003
5	DIS	.7829	.6129	.0090	.003

COL 10

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.6261	.3919	.3919	.001
2	REL 2	.6375	.4064	.0145	.004
3	DIS	.6469	.4185	.0121	.011
4	DIFF	.6726	.4524	.0339	.017
5	REL	.6844	.4685	.0161	.032

COL 11

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	DIS	.8751	.7658	.7658	.000
2	CON REL	.9300	.8649	.0992	.000
3	DIFF	.9332	.8709	.0060	.000
4	REL	.9339	.8723	.0014	.000
5	REL 2				

COL 12

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	DIFF	.3886	.1510	.1510	.067
2	REL	.4311	.1859	.0349	.128
3	CON REL	.4583	.2100	.0241	.204
4	TC	.4756	.2262	.0162	.302
5	REL 2	.4862	.2364	.0102	.420

TABLE 20

The summary table of the stepwise multiple regression analysis for one teacher. Order of inclusion of variable is determined by the respective contribution of each variable to explained variances.

STH 3

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	REL	.6151	.3783	.3783	.002
2	DIS	.6976	.4866	.1083	.001
3	REL 2	.7149	.5111	.0245	.003
4	DIFF	.7152	.5115	.0004	.009
5	CON REL	.7157	.5122	.0007	.022

STH 4

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	DIFF	.8718	.7601	.7601	.000
2	CON REL	.9067	.8220	.0619	.000
3	REL	.9144	.8360	.0140	.000
4	DIS	.9196	.8457	.0096	.000
5	REL 2	.9201	.8466	.0010	.000

STH 5

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.5749	.3305	.3305	.005
2	DIS	.6571	.4318	.4318	.005
3	REL 2	.6675	.4456	.4456	.012
4	REL	.6707	.4498	.4499	.030
5	DIFF	.6715	.4510	.4510	.064

STH 6

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.7148	.5109	.5110	.000
2	DIFF	.7715	.5952	.0842	.000
3	DIS	.2894	.6231	.0280	.000
4	REL	.8012	.6420	.0188	.000
5	REL 2	.8025	.6440	.0021	.001

TABLE 21

The summary table of the stepwise multiple regression analysis for one teacher. Order of inclusion of variable is determined by the respective contribution of each variable to explained variances.

MAN 1

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.8100	.6561	.6561	.000
2	REL	.8448	.7137	.0576	.000
3	REL 2	.8462	.7160	.0022	.000
4	DIFF	.8465	.7165	.0006	.000
5	DIS	.8466	.7167	.0002	.000

MAN 2

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.6609	.4367	.4367	.000
2	DIS	.7442	.5538	.1171	.000
3	REL 2	.7647	.5848	.0309	.000
4	DIFF	.7846	.6156	.0308	.001
5	REL	.7905	.6249	.0094	.002

BUC 1

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.6481	.4200	.4200	.001
2	DIS	.6758	.4567	.0367	.002
3	DIFF	.7026	.4937	.0369	.003
4	REL	.7235	.5234	.0298	.005
5					

BUC 2

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	DIS	.3866	.1494	.1494	.155
2	DIFF	.6337	.4015	.2521	.046
3	CON REL	.7541	.5686	.1671	.022
4	REL	.8045	.6471	.0785	.023
5	REL 2	.8049	.6479	.0065	.057

TABLE 22

The summary table of the stepwise multiple regression analysis for one teacher. Order of inclusion of variable is determined by the respective contribution of each variable to explained variances.

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.9198	.8461	.8461	.000
2	DIS	.9216	.8493	.0032	.000
3	DIFF	.9222	.8504	.0011	.000
4	REL				
5	REL 2				

STEP	VARIABLE ENTER	MULTIPLE R	R SQUARE	R SQUARE CHANGE	F
1	CON REL	.7606	.5635	.5635	.000
2	DIS	.7967	.6348	.0713	.000
3	DIFF	.8009	.6414	.0067	.000
4	REL				
4	REL 2				

accounts for most of the variance in Qual. This will be revealed in the stepwise regression analysis. When the effect of Con Rel is removed 90% of the partial coefficients between the overall rating and the remaining 4 independent variables are not significant.

The effect of Con Rel is also perceivable in the increment in R or R^2 due to the addition of a variable as a component attributable to that variable. Or in the reduction of R^2 to R square change when the variance due to Con Rel is removed from R^2. As shown in Tables 17 to 22, the values of R^2 are drastically reduced to R square change whenever Con Rel component is eliminated. Tables 10 to 15 also present regression coefficient, Bs. These coefficients are computed from unstandardized values. They indicate the amount of change in overall item rating accompanying a unit of change in the independent variables. An examination of the regression coefficient reveals that content relevance always has a positive relationship with overall item rating.

On the other hand, tables 17 to 22 show the results of the stepwise regression analysis. The order of inclusion of variables in the equation is determined by the criterion that the variable that explains the greatest amount of variance in the dependent variable enters into the regression equation first.

Content relevance (Con Rel) has entered into the regression equation first for 15 teachers out of 22. Item discrimination (Diss) has satisfied that statistical criterion for 3 teachers out of 22; item difficulty has satisfied the statistical criterion for 2 teachers out of 22; stability 1 (Rel) and stability 2 (Rel 2) each satisfied the statistical criterion for 1 teacher out of 22.

8. CONCLUSIONS

Teacher ratings of difficulty, discrimination and stability etc. of test items depend very much on teachers' perception of the item's content relevance to what has been taught in their classes. Most teachers' judgments of these item properties were not independent of the teachers' judgments of the item's content relevance.

The results of the stepwise regression analysis show that variable Con Rel, the content relevance, entered into the regression equation first 15 times out of 22. This indicates that variable Con Rel contributes more to the variance of Qual, the overall rating, in most of the cases than any other item property.

In both analyses Con Rel has more significant relationships with Qual than any other variable has with Qual. When Con Rel enters into the regression equation first, the relationships of other variables with Qual consequently improve. But when Con Rel component is removed from the variance only a few significant partial correlation coefficients remain. This means that Diff, Dis, Rel and Rel 2 have positive significant relationships with Qual mainly (91%) through Con Rel.

The results are similar to findings reported by Ryan (1968) and reviewed in Chapter 8.

CHAPTER 9

TESTS AND SCALES

The research so far suggests a difference in expectations and objectives between professional teachers and psychometricians with respect to the use of assessment procedures. In addition, the data suggest that teachers are not reliably able to estimate item properties. This part of the research confirms the work by Revinelli et al (1976) and the more recent claim in the T.E.S. (Seddon, 1982). From the evidence provided, teachers continually confused other item properties with content relevance. When presented with an item to make a judgment of such properties as item difficulty discrimination, reliability, etc., the teachers tend to make their judgments in accordance with the level of relevance that can be attached to the item. The question of trainability of teachers has not been looked at. It is probable that teachers can be trained to make adequate judgments. Scanty but supporting evidence comes from the study of Lorge and Diamond (1953). This view still needs to be empirically verified.

So far in this investigation we have concentrated on the reaction of the teachers to tests and assessments. In the final stage of the research we are concerned with the reactions of the students. In other words, it is necessary to investigate the consumers' verdict on the perceived validity, reliability, objectivity, fairness and acceptability of each of these two types of test.

In order to carry out this investigation, a professionally constructed test was compared with a psychometrically constructed test on criteria which incorporated both the professed objectives of the psychometricians and the professional teachers.

In order to carry out this investigation, a profesionally constructed test was compared with a psychometrically constructed test on criteria which

incorporated both the professed objectives of the psychometricians and the professsional teachers.

What is being investigated in this section is how tests constructed according to statistical or psychometric procedures and tests constructed according to professional teachers' judgments are differentially perceived by students from whom these tests are intended.

At this point one needs to define what is meant by a test, a statistical or a psychometric procedure of test construction, and a professional procedure of test construction. "A test is a set of tasks which is presented to the testee in a standard form and yields a numerical score or set of scores" (Annett, 1974). A statistical or psychometric procedure of test construction is that process in which test items have been written, tried out on a sample of subjects and then subjected to a statistical analysis, so that all test item properties such as difficulties, discrimination, validity, reliability etc., for each item, become known in advance. By professional procedure of test construction, we mean teacher-made tests which have not been subjected to statistical analysis, so that no test item property is known in advance. Two professionally constructed tests and two psychometrically constructed tests are included in this comparison.

1. THE PROBLEM OF COMPARISON

However, as mentioned earlier, the problem encountered when professional tests are compared with psychometric tests is that there is no independent criterion of comparison. Usually, the criteria available to judge between the two procedures are not independent of the two methods to be evaluated. The criteria are based either on human judgments or on statistical evidence. Each criterion is believed to be

biased in one way or another. Secondly, as shown below, the psychometric and the professional procedures of test constructions overlap a great deal in the process of test construction. Therefore, there is no such thing as a perfectly objective test. Psychometric tests are merely less subjective than professional tests.

2. THE TEST PLAN

The aim of this section is to expose the complexity of the tasks which the test constructor is required to engage in, in order to generate tests.

Proper planning of a test is the first and the most valuable step in test construction because all subsequent steps of the test development depend on the initial planning. A test is adequately planned when all the variables are accounted for. Reliability, validity and the usefulness of a test depend very much on how adequately the test has been planned. Probably the first point in test planning, one has to be clear on, is a statement of the purpose of the test to be constructed. One must have a clear idea of what is to be assessed. He must think about what he hopes an assessee will be able to do as a result of previous learning. The purpose for which the test is to be used determines the type and the property of test items to be constructed. If the test constructor has a vague conception of the purpose of the test, he will not be able to compose the relevant types of test items. So, one has to ask oneself in advance what he is testing for. Is he testing for placement; to monitor learning progress; to diagnose persistent learning difficulties; or to certify pupils at the end of a course? (Thorndike & Hagen, 1977; Wesman, 1971). What is expected of the test constructor is knowledge of the appropriate test construction procedure required for each decision-making situation. Clear intention of what is to be tested does

not only guide one to construct appropriate types of test items but also directs the process of instruction to be geared towards the desired goals.

For example, in testing for placement, one would be interested in determining, from the beginning of the instructional programme, the amount of skills so far achieved by the assessees so that they can be properly allocated to instructional programmes. Information obtained from placement tests aid educators to carry out instruction at the appropriate level. Imparting instruction at the appropriate level helps one avoid teaching pupils too high above their level of understanding for to do so is to frustrate and discourage the learner. To teach him too low below his understanding is to make teaching boring to the learner. Placement decisions are not only confined to the entry level of performance. They are also made to determine as to which option a particular student has to pursue. The purpose of placement tests is to determine whether particular persons have the necessary requisite skills for a programme of instruction. Placement tests have a limited area of content and relatively low level of difficulty items. However, measures of final achievement can as well serve the purposes of placement. Decisions about the entry level and allocation of individuals into appropriate programmes of instruction can be made on the basis of information obtained from final achievement tests.

In testing for monitoring academic progress, on the other hand, the test constructor must be aware that tests designed to monitor learning progress require different approaches of construction from tests designed to aid placement decisions; to certify students; or diagnose persistent learning difficulties. Since formative tests are intended to correct weakness and improve learning, the items in formative tests are typically criterion-referenced test items. To provide continuous feedback, formative

tests are frequently given to measure pupils' mastery of specific contents.

It is said that "the diagnostic test takes up where the formative test leaves off" (Gronlund, 1981). If the prescriptive treatments decided on the basis of the information obtained from formative testing are not corrective enough, the next step one has to take is to design a diagnostic test. The purpose of the diagnostic test is to probe the source of the persistent learning difficulties. The diagnostic test should be composed of relatively easy items from each specific area of the content, and constructed in such a way that all the errors made by the assessees can be tapped.

The summative tests given at the end of a course of instruction are mainly intended for certification. These tests are typically norm-referenced tests,with items of varying degrees of difficulty.

The second step the test constructor ought to take is to develop the test specifications. Test specifications are probably the only priori assurance one can have that the test in question is a valid measure of the instructional objectives (Millman, 1980). The purpose of the test specifications is to define the scope and emphasis of the test and to constrain the test constructor so that he produces balanced test items. There are several ways of devising test specifications. Some of these are described in Popham (1978b; 1980). A well known device of test specification strategy is the content by objectives table which relates the instructional objectives to the subject matter content (Thorndike & Hagen, 1977; Gronlund, 1981). One advantage of the test specifications is that it serves as a useful guide in terms of the objectives and in terms of the course content. All learning outcomes receive appropriate emphasis when appropriate test specifications are

adhered to. With the help of the table of test specifications, the test constructor must decide the emphasis or the proportion of test items across instructional objectives and the content areas. The actual construction of the appropriate test items will measure whether or not the respondent possesses the behaviour in question. The test constructor must also decide what to do with behaviours which may not be measured by written tests.

Item types can be classified in different ways: One way is in terms of objective test items, essay questions, performance test items, etc. In this classification,item types are not necessarily confined to measure one type of learning outcome. Objective multiple-choice test items, for example, can be used to measure knowledge, understanding, application etc. Another way of classifying test items is in terms of the behavioural objectives/the items (Bloom, 1956).

Knowledge and due consideration on the part of the test constructor of the match between test items and the behaviours to be measured are essential to the construction of valid and reliable tests.

The point we are making is that the consequence of the failure to select the appropriate item types for the behavioural objectives is a mismatch between test items and the behaviours to be tapped. This mismatch lowers the content validity of the test and undermines all the previous steps of the test plan.

2.1. SUMMARY

The section on test plan discussed the extent to which all the desirable test qualities depend on the adequacy of the test plan; the adequacy with which the test plan can be executed depends on the clarity of the purpose of the test, i.e. the type of decision to be made

on the information provided by the test; a clear conception of the expected abilities of the persons to be tested; the ability with which the test maker delimits and strictly follows a domain specification which constrains him to remain within the domain to be sampled and guides him to produce balanced test items. From there the test constructor takes the appropriate steps to construct a psychometric or a professional test.

3. THE COMMON STEPS IN TEST CONSTRUCTION

The most common steps taken in both the psychometric and the professional procedures to construct a test are shown here. The steps enumerated under the professional procedure of test construction are mainly applicable to professionally constructed tests. However, most of the steps taken to construct professionally constructed tests also apply to other assessment techniques such as ratings, oral tests, essay examinations etc. The most common steps taken in the psychometric approach to test construction are the following:

1. Identifying the domain of universe from which the test is to be sampled.
2. Objectives and learning outcomes are identified and defined.
3. The subject matter content is outlined.
4. A table of specifications which relates objectives to subject matter is developed.
5. The test constructor samples from the domain of universe.
6. The content validity of the items sampled from the domain is based on rational judgment.
7. The property of the subject to be measured is conceptualized.
8. The adequacy of instructions for administering directions for the pupils, time limit and scoring procedures are determined.

9. The test is tried out experimentally on a sample of subjects.
10. Test scores are subjected to statistical analysis.
11. Bias and ambiguity are eliminated.
12. Difficulty, discriminating power, validity and reliability are determined.
13. Test items are correlated with the total score, or other methods are used.
14. Selection and rejection of test items are made in terms of their desired characteristics.
15. The adequacy of instructions for administering, directions for pupils, time limit and scoring procedures are determined.
16. Finally, the test is administered to the target population.

The most common steps taken in the professional approach to test construction are the following:

(A)1. Identifying the domain of universe from which the test is to be sampled.
2. Objectives and learning outcomes are identified and defined.
3. The subject matter content is outlined.
4. A table of specifications which relates objectives to subject matter is developed.
5. The test constructor samples from the domain of universe.
6. The content validity of the items sampled from the domain is based on rational judgment.
7. The property of the subject to be measured is conceptualized.
8. Test items are assembled and reviewed by the test constructor.
9. Some items are selected by subjective judgment.
10. The adequacy of instructions for administering, directions for the pupils, time limit and scoring procedures are determined.

11. Finally, the test is administered to the target population.

(B) Without written tests, the teacher may, according to his knowledge of the subjects, use personal judgment to assign grades to his students; or the teacher may give oral examinations; he may give his students essay questions etc.

4. METHODOLOGY

Two types of comparison were planned. The first was made on the basis of student scores on each of the two types of test. The second was made in terms of students' judgments on 5 point scales. These scales assessed the degree to which each type of test took account of accuracy, prior knowledge of the examinees, subject matter, objectivity and care in construction.

4.1. Subjects

Seventy O'level students acted as subjects for this part of the investigation. They were 32 males and 38 females. They were in the fifth form about to enter O'level examinations in biology and other subjects. They came from two schools.

4.2. The Instrument

4.2.1. (a) Constructing the Tests

Two of the tests were constructed by teachers who had experiences in constructing multiple choice test items. The two other tests were psychometric tests from the same discipline and standardized on a population similar to that which performed the tests.

The methods of test construction represent different experimental treatment effects. To obtain a valid comparison both methods should be applied to the same content.

The material for the two psychometric tests was obtained from one of the established school examination boards of England. Items dealing with biology and assessing the currently taught syllabus were chosen from the item banks. Item psychometric characteristics had previously been computed from an operational sample of 30,692 candidates (1981) and 1025152 candidates (1982). These characteristics were difficulty level, discrimination and the proportion of students who passed the test. The items were chosen to ensure an adequate coverage of the total syllabus. To obtain equivalent forms of the same test two sets of 15 items were selected from the bank. The items were matched in terms of these three item properties.

To obtain the professional tests the Examination Board provided items as they had been submitted by the teachers before they had been screened by the test experts and pre-tested for standardization. Again two sets of 15 items were selected. The items were balanced for syllabus coverage and content. Two professional and two psychometric tests were the outcome. All four tests were balanced for content and syllabus coverage.

The procedure used by the Examination Board were as follows: Test items were drafted by experienced teachers and reviewed by test experts. The teachers selected are given training in writing multiple choice items before they are commissioned to draft the test items. The item reviewers have already been chosen for their particular skills in testing and test construction. The tests presented by the teachers and not yet modified by experts or subjected to empirical refinements,are

in this study as professional tests.

The procedure followed to develop psychometric tests involved the individual item writers who have special training as above. Items drafted were passed to a second panel who reviewed and edited them. Item reviewers accepted some of the items and rejected others on the bases of personal judgments. Items accepted were pre-tested on a representative sample of subjects of not less than 250. The scores made by the subjects on the test were analysed and the relevant psychometric indices were computed for each item. Items which had biserial of greater than .20 and less than .90; δ values between 9 and 17; p values between .20 and .80 were retained to form the psychometric tests.

Instructions and advanced information given to item writers are shown in Appendix III . The actual items used were considered confidential by the Examination Board and are not presented in this thesis.

4.2.2. (b) Constructing the Judgmental Scales

The second part of the measuring instrument was a scale of 15 criteria on which the judges evaluated the tests. A scale was constructed to evaluate the tests on three factors identified in the first stage of the study. The scale consisted of 15 items. In addition, two open-ended questions were inviting an overall judgment and general comments. The scale was scored on 5 points. Students' judgments were invited on the tests as a whole. Thus, the student after taking the examination had to fill in the scales recording his reaction to the examinations he had taken. Each student completed two examinations and two scales; one was a professional examination and the other was psychometric.

On the bases of the criteria obtained from teachers and other pro-

essionals' responses to the attitude statements, the scale for judging the relative worth of the tests was constructed. In the first stage we obtained evidence of three principal dimension teachers' use in judging the effectiveness of tests for their pupils. The interest in this third stage is to ascertain which of the two types of test is perceived as preferable to students. The criteria of preference were taken from the three dimensions in the first stage. In other words, we want to find out which of the two tests, each constructed to a different method, manifests more of the qualities believed to have been desired by teachers and expressed through their responses to the attitude statements.

4.3. THE PROCEDURE

It was too much to ask any one subject to judge or perform all the 4 tests, and to read the list of the criteria of comparison. To reduce that load, only one psychometric and one professional test was allocated to each judge or student. Each two tests and the criteria for comparison accompanied by instructions were presented to the judges at the same time. The problem of overcoming the cumulative progressive effects on the judgments and on students' performance was dealt with by presenting the tests to the subjects in counter-balanced order. The students received the tests in the orders shown below. The tests were administered in groups by the students' class teachers as mock examinations. To evaluate the tests on the criteria given, subjects were given the following instructions:

INSTRUCTIONS

There are two multiple choice biology tests. The two tests were constructed by two different teachers. You are asked to read the tests in the order they are given to you. Please read the test items carefully and answer all the questions in the test. Then, judge the quality of the test by filling the 5-point scale attached to each test. THANK YOU.

TESTS X SUBJECTS

Order of presentation of tests to subjects

SUBJECTS	TESTS	SUBJECTS	TESTS
S1,S17,S33	PR1 + PS1	S9,S25,S41	PS2 + PR1
S2,S18,S34	PS2 + PR2	S10,S26,S42	PR2 + PS1
S3,S19,S35	PR2 + PS2	S11,S27,S43	PS1 + PR2
S4,S20,S36	PS1 + PR1	S12,S28,S44	PR1 + PS2
S5,S21,S37	PR1 + PS1	S13,S29,S45	PS1 + PR1
S6,S22,S38	PS1 + PR2	S14,S30,S46	PR2 + PS2
S7,S23,S39	PR2 + PS1	S15,S31,S47	PS2 + PR2
S8,S24,S40	PS2 + PR1	S16,S32,S48	PR1 + PS1

PR1 = PROFESSIONAL 1

PR2 = PROFESSIONAL 2

PS1 = PSYCHOMETRIC 1

PS2 = PSYCHOMETRIC 2

4.4. THE CRITERIA FOR JUDGMENT FOR THE STUDENTS

CRITERIA	Very or very much	Quite	Fairly	Somewhat	Not at all
1. How easy is it to judge the level of difficulty of this test?					
2. How accurate is the test to measure your ability in biology?					
3. How accurate is the test in predicting how well you would do in a new area of biology?					
4. Some tests do not allow a student to show his/her true knowledge; how appropriate is the test to measure your true knowledge in this subject?					
5. How accurate is the test in sorting out those who would pass A-Levels?					
6. How much do you say the person who constructed this test has taken into account your prior knowledge of biology?					
7. With what degree of confidence would you have accepted grades you receive in this test?					
8. How fair would it be to rank pupils in your class on the basis of scores they have made on this test?					
9. If you had the choice would you. prefer another type of test to give you a fairer chance?					

CRITERIA	Very or very much	Quite	Fairly	Somewhat	Not at all
10. From what you have seen of this test how confident would you feel that the test was properly and carefully made up?					
11. How objective is the test as a measure of knowledge in biology?					
12. How proper is it for a teacher to compare pupils on scores they have made on this test?					
13. How fair is it to use this test to tell how much you know in biology?					
14. To what extent do you believe this test has been made according to a good and sensible plan?					
15. To what extent do you think that the questions in the test were properly checked before you took the test?					

16. How well do you think you did in this test? Give yourself mark out of 10 _____.

17. Say what you feel was wrong with this test. Say what you feel was good about this test.

(WRITE YOUR COMMENTS ON THE BACK OF THE SHEET)

18, Which of the two tests you have taken is better? Why?

THANK YOU

CHAPTER 10

RESULTS AND ANALYSES

The two psychometric tests and the two professional tests are evaluated and compared on the evidence obtained from the following:

(a) Students' performance on the tests and

(b) Students' ratings of the qualities of each test.

The responses are evaluated on the indices listed below.

(a) Internal Consistency

(b) Reliability

(c) Construct Validity

(d) Perceived Accuracy

(e) Perceived Fairness to Examinees

(f) Perceived Objectivity

Before we present the results of the analysis we shall briefly recapitulate some descriptions of the procedures and reasons for proposing these particular procedures to be the appropriate ones that could provide more relevant evidence on which the tests are to be evaluated and compared.

1. QUANTITATIVE ANALYSES

1.1. Internal Consistency

The present tests are performed by students only once. The main source of variation in the tests and the one we are particularly interested in is that which is due to content sampling.

The tests were intended to measure only one characteristic, the characteristic in question being the pupil's ability in biology. Internal consistency is more appropriate to estimate the reliability of single

administration tests (Burns and Dobson, 1981). The procedure is also more appropriate to estimate variation which is due to content sampling. Internal consistency is also a measure of homogeneity (Anastasi, 1976). Each test is evaluated and compared with other tests to the extent that all the items in the test measure the same characteristic. However, we recognise the fact that test items rarely measure only one thing and nothing else. Items in a test always measure more than one thing. It is also true that some tests are closer to being unifactor measures than other tests. Therefore, the present tests will be evaluated and compared with each other to the extent that each test approximates unidimensionality.

A unidimensional test like biology is accurate to the extent to which all its items measure the same thing. Internal consistency reliability estimate measures the extent to which all items in a test measure a common attribute (Jensen, 1980). This procedure is said to provide a better estimate of test reliability in most cases (Nunnally, 1978). The formulae commonly used to estimate the reliability coefficient of the test are KR20 and Cronbach's Coefficient α. The first is commonly used when items in the test are scored dichotomously and the second is commonly used in multipoint item tests. The internal consistency reliability index is low when the interitem correlations are low. This situation in turn indicates either a lack of homogeneity or the diversity of what the items measure. If the data is in dichotomous form, α is equivalent to the reliability coefficient KR20 (Hull and Nie, 1981).

1.2. RELIABILITY

Another procedure proposed to estimate test reliability is the analysis of variance. Reliability is defined as the ratio of true variance to the total score variance. Reliability can be computed from

one of these alternative formulae: $V_H = \frac{Vt}{Vtot}$, $r_H = 1 - \frac{Ve}{Vtot}$, $r_H = \frac{Vtot-Ve}{Vto}$. We intend to use the formulae that will be more convenient. Analysis of variance is a method of breaking down the total variation yielded by the measuring instrument into component sources of variance (Kerlinger, 1973). The partitioning of the variance into error and systematic variances is achieved by the analysis of variance procedure. The rationale of using Anova pertains to obtaining these variance components of the tests. The relative sizes of the components provide relevant information for evaluating test reliability. The two components pitted against each other for each test are the true variance and the error variance. Each test is evaluated on the magnitudes of these two components. The smaller the error component, and the greater the true variance, the more accurate is the test.

A second index of reliability is the Error of Measurement. This is only to confirm the reliability indices already obtained for the tests. Standard Error of Measurement does not contradict other reliability procedures. Standard Error of Measurement and Reliability Coefficient are alternative ways of expressing test reliability.

When test items measure irrelevant variables plus relevant ones, the error of variance is greater. When they measure only relevant trait(s), the error of variance is smaller. The question to be asked is which of the 4 tests investigated in this part of the study has more accurate scores; which test has a larger error of variance or, which test has a larger zone of uncertainty along the scale continuum? The accuracy (reliability) of a test can be expressed in terms of standard error of measurement. The procedure measures the degree of accuracy of test scores. The magnitude of the standard error of measurement indicates the degree of accuracy of a test.

1.3. CONSTRUCT VALIDITY

The factor analysis procedure was proposed in order to investigate the tests' construct validities. The tests used in this section of the study were designed to measure the same attribute of the pupils who performed the tests. The relevant attribute was pupils' ability in biology. The tests were evaluated and compared to the extent that each test accurately measures pupils' ability in biology. The tests were evaluated and compared on their relative loadings on that attribute. The extent to which the test's variance is accounted for by the relevant factors shows the test has a factorial validity. The higher the proportion of the total variance accounted for by the relevant factor the more valid and relevant is the factor analysis to examine the construct and content validities of measures. In construct validation, factor analysis is used to test statistically the adequacy of the factorial composition of the measure. With factor analysis one determines whether or not an expected internal structure of a particular measure exists. Some tests are constructed to measure one attribute. Others are constructed to measure several factors. When unidimensionality is at issue, the test is factorially valid which with fewer factors accounts for a greater portion of the test's total variance. On the other hand, when multidimensionality is desired, the test is factorially valid to the extent to which the factor analysis confirms the real dimensions in the measure. For example if a test measuring 4 difficult attributes is factor analysed, one should obtain 4 different clusters of variables. However, one problem with factor analysis is that, in addition to the desired factors, the tests also measure some irrelevant factors.

The present tests were designed to measure a single academic discipline and can be seen as unidimensional. The characteristic of the pupils measured by the tests was their ability in biology.

2. THE RESULTS OF THE ANALYSES

2.1. Reliability Analyses

Two reliability analyses were carried by the computer programme 'RELIABILITY'. Both the analysis of variance and the Alpha were printed simultaneously. KR20 commonly used with data in dichotomous form was not available in the computer programme. However, according to the author of the programme, if the data is in dichotomous form, Alpha is equivalent to the reliability Coefficient KR20 (Hull and Nie, 1981). Therefore, Model Alpha and the analysis of variance were requested to be printed for each test. After the source of variation, ss,df, and the means of squares are given, reliability can be estimated by the following formula: $V_{tt} = \frac{Ve}{Vtot}$ (Burroughs, 1975). The results of the analysis of variance are shown in Tables 1 to 4.

TABLE 1

Table of Analysis of Variance for Test: PR2

Source of Variation	SS	DF	MS	F	Sig.	Rel. Coeff.
Between people	14.74000	39	.37795			
Within people	128.0000	560	.22857			
Between measures	29.09000	14	2.07786	11.47012	.0001	
Residual	98.91000	546	.18115			.52070
Nonadditivity	1.131633	1	1.31633	7.35087	.0069	
Balance	97.59367	545	.17907			
TOTAL	142.7400	599	.23830			

TABLE 2

Table of Analysis of Variance for Test: PR1

Source of variation	SS	DF	MS	F	Sig.	Rel.Coeff.
Between people	7.91351	36	.21982			
Within people	122.13333	518	.23578			
Between measures	28.42523	14	2.03037	10.92017	.0001	
Residual	93.70811	504	.18593			.15417
Non addivity	3.33533	1	3.33533	18.56390	.0001	
Balance	90.37278	503	.17967			
TOTAL	130.04685	554	.23474			

TABLE 3

Table of Analysis of Variance for Test: PS2

Source of variation	SS	DF	MS	F	Sig.	Rel. Coeff.
Between people	14.15726	38	.37256			
Within people	131.73333	546	.24127			
Between measures	31.27521	14	2.23394	11.83038	.0001	
Residual	100.45812	531	.18883	3.49276		.4932
Non addivity	.65647	1	.65647		.0622	
Balance	99.80165	531	.18795			
TOTAL	145.89060	584	.24981			

TABLE 4

Table of Analysis of Variance for Test: PS1

Source of Variation	SS	DF	MS	F	Sig.	Rel.Coeff.
Between people	17.80180	36	.49449			
Within people	114.13333	518	.22033			
Between measures	23.44865	14	1.67490	9.30864	.0001	
Residual	90.68468	504	.17993	5.52797		.6361302
Non additivity	.98579	1	.98579		.0191	
Balance	89.69889	503	.17833			
TOTAL	131.93514	554	.23815			

To be able to interpret the significance of r from zero or the significance of the difference between rs, one needs to know the amount of error of measurement associated with each r. Standard errors associated with the reliability coefficients vary with the size of the samples and with the sizes of the rs. In this case, transformation of the rs to the corresponding zs (Fisher's z) is invoked (Burroughs, 1975; Ferguson, 1976). The sample sizes and the standard errors associated with each test are shown in Table 5.

TABLE 5

Table of Internal Consistency Reliabilty Indices, Standard Errors associated with the rs, corresponding Fisher's zs and the significant levels of the rs.

Tests	Sample Size	Alpha	Z	Standard Error associated with rs	Sig. r
Psychometric 1	37	.63613	.76	.1715	.01
Psychometric 2	39	.49315	.55	.1664	.01
Professional 1	37	.15418	.156	.1715	NS
Professional 2	40	.52069	.58	.1643	.01

The difference between any two zs to be significant, the ratio of the obtained difference between zs and standard error of the difference of z must be equal or greater than 2 times the se of the difference of z (Burroughs, 1975). For example, the standard error (se) for the psychometric test 1 is $\frac{1}{\sqrt{37-3}} = \frac{1}{\sqrt{34}}$. $se^2 = \frac{1}{34} = 0.0294$. The standard error for the professional test 1, which is 0.0294, is similarly obtained. se of difference of z = 0.0294 + 0.0294 = 0.0588 .242. The difference between zs associated with psychometric test 1 and the professional test 1 is .6 (.76 - .156). The ratio in question is 2.52 ($\frac{.61}{.242}$) which is greater than twice the standard error of the difference of z. Therefore, the difference between z of .76 and z of .156 is significant. (.02). The differences between ts associated with other tests are not significant. All the rs are significant at .01 except that associated with the professional test 1. As shown in tables 1 to 5, the reliability coefficients obtained for each test by the two reliability methods are identical. The error of measurement associated with the rs (see Table 5) are similar for all 4 tests.

The results of the reliability analyses indicated that there were no overall differences between the reliability indices computed for the two professional and the two psychometric tests. There were 4 possible comparisons between the two sets of tests. Of the 4 possible comparisons, a significant difference between the reliability indices was obtained for a single case. The reliability index computed for psychometric test 1 was significantly different from the reliability index computed for professional test 1. The error of measurements associated with the tests (see Table 5) were similar for all tests. From these results, it was concluded that there were no overall significant differences between the psychometrically constructed tests and the professionally constructed tests.

THE RESULTS OF THE FACTOR ANALYSIS

As shown in tables 6 to 9, each test was loaded on several factors. The sizes of the factors and the structures are similar in all 4 tests The number of factors with the eigenvalues of greater than 1 and the percentage of the variance the factors accounted for are also similar. As shown in tables 6 and 7, the number of factors with the eigen values of greater than 1 are 6 and 7 for professional tests 1 and 2 respectively. The number of factors with the eigen values of greater than 1 are 5 and 6, for the psychometric tests 1 and 2, respectively. On the other hand, the percentages of the variances accounted for by these factors are 68.6 for professional test 1 and 74.0 for professional test 2. For the psychometric tests, the percentages are 65.2 for test 1 and 68.6 for test 2. The two sets of tests are not significantly different in terms of the number of factors with the eigenvalues of greater than 1 or in terms of the percentages of variances the factors accounted for.

However, it is difficult to compare the factorial validities of tests with multiple factors. Each of the tests above has several indices to be compared with several other indices from each of the other two tests.

One may view validity from variance breakdown. A test's total variance can be partitioned into its components. These components are the common factor variance specific variance and the error variance.

TABLE 6

Table of the Results of the Factor Analysis for PR1

VAR	Communality	Factor	Eigen value	PCT of VAR	CUM PCT
1	.7027	1	2.33971	15.6	15.6
2	.1081	2	2.08795	13.9	29.5
3	.7568	3	1.84592	12.3	41.8
4	.6486	4	1.72812	11.5	53.5
5	.2162	5	1.20617	8.0	61.4
6	.1351	6	1.08741	7.2	68.6
7	.4324	7	.89443	6.0	74.6
8	.5135	8	.81604	5.4	80.0
9	.1622	9	.74656	5.0	85.0
10	.4595	10	.67248	4.5	89.5
11	.1622	11	.46203	3.1	92.6
12	.1081	12	.36436	2.4	95.0
13	.1622	13	.32415	2.2	97.2
14	.5676	14	.23312	1.6	98.7
15	.4865	15	.19155	1.3	100.0

TABLE 7

Table of the Results of the Factor Analysis for PR2

VAR	Communality	Factor	Eigen Value	PCT of VAR	CUM PCT
1	.44312	1	2.66462	17.8	17.8
2	.22272	2	2.07278	13.8	31.6
3	.58037	3	1.45166	9.7	41.3
4	.35523	4	1.36124	9.1	50.3
5	.44040	5	1.23623	8.2	58.6
6	.25330	6	1.21119	8.1	66.7
7	.21817	7	1.10128	7.3	74.0
8	.28513	8	.90484	6.0	80.0
9	.63687	9	.72368	4.8	84.9
10	.52761	10	.54732	3.6	88.5
11	.51173	11	.52598	3.5	92.0
12	.31666	12	.47168	3.1	95.1
13	.45887	13	.29616	2.0	97.1
14	.38181	14	.25795	1.7	98.8
15	.34145	15	.17340	1.2	100.0

TABLE 8

Table of the Results of the Factor Analysis for PS1

VAR	Communality	Factor	Eigen Value	PCT of VAR	CUM PCT
1	.52139	1	2.94685	19.6	19.6
2	.35214	2	2.12025	14.1	33.8
3	.41576	3	1.89481	12.6	46.4
4	.35719	4	1.57087	10.5	56.9
5	.51581	5	1.24960	8.3	65.2
6	.54016	6	.96600	6.4	71.7
7	.57913	7	.84311	5.6	77.3
8	.47037	8	.74954	5.0	83.3
9	.63416	9	.66222	4.4	86.7
10	.35131	10	.61373	4.1	90.8
11	.45957	11	.39577	2.6	93.4
12	.51292	12	.36845	2.5	95.9
13	.48739	13	.27984	1.9	97.7
14	.61870	14	.18610	1.2	99.0
15	.51081	15	.15285	1.0	100.0

TABLE 9

Table of the Results of the Factor Analysis for PS2

VAR	Communality	Factor	Eigen Value	PCT of VAR	CUM PCT
1	.42978	1	2.79502	18.6	18.6
2	.59174	2	2.14761	14.3	33.0
3	.62005	3	1.67722	11.2	44.1
4	.39444	4	1.51247	10.1	54.2
5	.42074	5	1.11880	7.5	71.7
6	.45362	6	1.04601	7.0	68.6
7	.36339	7	.99569	6.6	75.3
8	.43983	8	.85094	5.7	81.0
9	.27943	9	.77241	5.1	86.1
10	.58523	10	.56637	3.8	89.9
11	.56262	11	.49183	3.3	73.2
12	.57580	12	.38546	2.6	95.7
13	.45569	13	.29281	2.0	97.7
14	.39532	14	.21772	1.5	99.1
15	.54286	15	.12964	.9	100.0

TABLE 10

The sum and the mean of the communalities for the 4 tests and the number of items in each test

Tests	No. of items	Sum of communality	$\bar{x}$
Professional Test 1	15	5.6217	.37478
Professional Test 2	15	5.96984	.39799
Psychometric Test 1	15	7.32681	.48846
Psychometric Test 2	15	7.11054	.474036

The validity of a variable in a measure, the test item in this case, is the portion of the total variance the variable shares with other variables in the measure. The portion of the variance shared is called communality.

Now, to obtain a single validity index for each test, the communalities of each test may be added and divided by the number of items in the test. The 4 tests are then compared on the 4 indices obtained. Table 10 shows the sums and the means of the communalities for each test. However, the interpretation and the subsequent comparisons of the tests in terms of the sizes of their communality means is valid when it is assumed that all the common factors measure relevant aspects of the pupils' ability in biology.

A t-test for independent samples was computed between the highest and the lowest means in Table 10. The result of the t-test showed that there is no significant difference between the highest and the lowest means. Therefore, there is no overall significant difference between the two types of tests in terms of the sizes of their comm-unalities.

3. QUALITATIVE EVALUATION OF THE TESTS

All the decisions so far made about the relative worth of psychometric vs professional approaches to tests and test construction, were based on evidence obtained from both qualitative and quantitative evaluations of the methods compared. In this section too, the qualities of the tests were evaluated directly on a set of criteria. The identities of the tests were not revealed to the judges. The purpose of obtaining the subjective rating was to determine whether the subjects expressed more favourable responses toward the psychometric test or toward the professional test. There were 66 students. Each student performed both tests and rated each test on the criteria given. The five-point scale was constructed in such a way that high ratings corresponded to favourable responses. The ratings made by the students on each type of test were compared. Since there was a single group of subjects under two conditions, a t-test for correlated samples has been computed. The result of the analysis of the t-test is shown in Table 11.

TABLE 11

T-test of the difference between two means for correlated samples

N	ED	ED^2	T	SIG
66	317	18461	2.42	.02

A separate analysis was done for each subscore, dealing with the 3 sections of the scales, i.e. accuracy, fairness, objectivity.

Each 5 items derived from a different factor have been scored separately. As shown in Table 12, the lowest total scores and the lowest total subscores were obtained for the professional test 1. The 3 subscores within each test were not significantly different from each other.

TABLE 12

The 3 subscores made by each student and their totals

A	B	C	A	B	C	A	B	C	A	B	C
14	19	16	13	18	15	9	8	6	10	11	6
10	15	12	15	18	16	12	6	7	11	9	11
11	7	8	16	11	15	8	7	8	15	11	15
18	22	21	13	10	14	13	9	11	9	9	9
17	20	20	15	11	9	8	10	8	11	13	18
17	18	18	19	15	16	7	12	6	15	18	16
12	15	9	14	10	14	7	6	6	12	9	9
14	13	12	18	18	16	9	6	10	11	13	13
15	16	13	14	15	13	11	6	8	14	16	12
11	14	14	15	11	16	7	8	5	14	21	12
15	17	21	16	18	18	18	11	14	17	21	19
19	18	18	13	10	9	8	10	8	16	17	11
9	5	6	14	11	13	16	11	18	16	20	20
11	5	7	9	5	6	8	10	8	15	13	13
19	15	20	21	18	19	14	10	14	15	18	21
11	8	8	11	7	8	14	14	19	11	13	17
9	7	13	14	19	16	10	9	11	12	11	12
13	15	12	12	15	14	14	13	12	15	10	12
10	7	6	15	20	20	15	15	13	18	11	14
20	9	14	12	15	14	15	17	18	19	13	16
11	8	13	17	13	12	18	16	21	13	16	12
11	8	8	15	18	16	16	22	17	12	9	9
9	7	8	13	11	20	11	10	9	6	5	6
12	6	10	15	20	20	7	6	5	12	11	7
9	11	8	15	17	18	9	6	6	11	12	14
14	14	7	16	17	12	16	12	15	14	15	9
11	8	9	18	18	16	11	12	10	7	5	7
5	5	5	7	7	7	8	9	11	10	10	11
11	8	10	7	9	11	10	8	6	20	9	14
13	11	14	16	13	17	17	12	17	15	9	9
10	15	12	12	12	12	14	14	9	10	8	12
Ex 386	366	372	480	470	474	367	335	348	443	415	419
Ex^2 5150	5074	5134	7066	7044	7084	4623	3933	4426	6099	5649	5651
	PS1 N=31			PS2 N=34			PR1 N=32			PR2 N=34	

TABLE 13

Tests and their Subscores

Subscores	PS1	PS2	$\bar{x}$	PR1	PR2	$\bar{x}$	$\bar{x}$
A	386	480	433	367	443	405	419
B	366	470	418	335	415	375	397
C	372	475	423	348	419	384	404
$\bar{x}$	375	475		350	426		

A = accuracy dimension
B = fairness
C = objectivity dimension

The students' ratings of the test's qualities reflect their performance of the tests. The most striking similarity between the students' ratings of the tests' qualities and the qualitative analysis of their performance of the tests is that the professional test 1 is rated by the students as the poorest test. The quantitative analysis of the performance of the students in the tests revealed that the professional test 1 has the poorest psychometric indices. The second similarity between the students' ratings of the tests' qualities and their actual performance on the tests is that the differences in qualities between the other tests is not great. The empirical analyses confirmed the students' ratings of the tests.

For example, the reliability coefficients of the psychometric tests 1,2 and the professional test 2 are .636, .4932 and .520 respectively,

while the reliability coefficient of professional test 1 is .154, which is significantly different from the other reliability coefficients. These similarities between students' ratings and their performance on the tests is an evidence that students ratings were valid and that students' judgments can be used to evaluate test qualities. The rank order correlation coefficient between ratings and the values obtained empirically was .80 with degrees of freedom of 4.

Another aspect of the qualitative valuation of the tests by the students concerned students' estimates of their performances on the text. Students were asked to estimate how well they thought they would do on the tests. Not all the students answered the question. However, the responses of those who answered the question are as follows: 26 students rated themselves on professional test 1, 24 on professional test 2, 23 on psychometric test 1, and 27 on psychometric test 2. On the other hand, the corresponding numbers who performed the test were: 37 students on professional test 1, 40 on professional test 2, 37 on psychometric test 1, and 39 on psychometric test 2. The actual means and the estimated means are shown in Table 14.

TABLE 14

Actual and Estimated Means of the Tests in Part 4

	Professional test 1	Professional test 2	Psychometric test 1	Psychometric test 2
Actual $\bar{x}$	37	43	39	48
Estimated $\bar{x}$	54	61	59	58

The estimated means are significantly larger than the actual means.

The students overestimated their abilities. They predicted the poorest performance on professional test 1 which confirmed the empirical analysis. However, Spearman's rank correlation coefficient between the actual means and the estimated means was .60 which is not significant.

The students were also asked to indicate which of the two types of test they have performed was better. The data obtained from the students' reactions to that question has very little to offer. First, the students made irrelevant comments about tests in general. Secondly, they equated the test's qualities with its difficulty level. Students thought more difficult tests measure students' ability better than easier ones. Others preferred easy tests. However, those who confined their responses to the relevant question have not shown any overwhelming preference for either type of test. The ratio was about 7 to 8, 8 for the psychometric tests and 7 for the professional tests. Also, the general comments students made about the tests had no distinct pattern. 45 of the 66 students who responded to item 9 in the scale preferred some other way of being assessed. However, none of them suggested any specific alternative methods of assessment.

4. CONCLUSION

The question to be answered in this part of the study was which of the two types of test ,each constructed according to a different method of test construction, was more accurate, valid and fair. The two types of test were compared on the evidence obtained from both quantitative and qualitative evaluations. From the stand point of students reactions to the tests, both quantitatively and qualitatively, psychometrically constructed tests were not better than professionally constructed tests in any of

the test qualities investigated in this study.

This finding contradicts the belief that psychometrically constructed tests have better content relevance, reliability etc. (Ausubel and Robinson, 1969; Anastasi, 1969). On the other hand, the findings support Rovinelli and Hambleton's (1976) arguments that systematic human judgments can be enough for the ensemble of suitable test items. They said any benefits obtained through statistical application to test items would not be commensurable with the extra labour involved. Similar arguments have been made by Ebel (1956) and Wesman (1971). With elaborate test specifications Ebel (1962) has shown that equivalent tests of high psychometric properties could be constructed by subjective judgments.

CHAPTER 11

CONCLUSION

The old controversy on the subjectivity of individual observations seems to have re-emerged in current educational decision-making. Its presence is seen in current attempts to give greater credence and prominence to headmasters' and headmistresses' assessment of pupils. The urge to flee from subjectivity seems to be submerged beneath a stronger urge to individualise decisions.

The position was put as if objective methods of measurement fail to take adequate cognisance of the individuality of the person being measured. Classical measurement theory caught up with this educational discontent and introduced the concept of criterion-referenced testing. The hope was that by referring an individual's performance to his own previous performance as criterion, individuality was respected. The contrast with the more traditional norm-referenced approach was obvious and welcomed by educators.

What remained unanswered was the point of immediacy and professional involvement. Teachers were asking not only for cognisance of the individual but appreciation of the cumulative knowledge that comes with years of experience and interacting with pupils. Where, they wanted to know, was the detailed knowledge of the individual pupil's personality being given its due place?

The research literature shows little serious attempt either to demonstrate the fallibility of professional teachers' reliance on teachers' judgment, or to indicate where such judgment is valid and where not. This is the starting point of this study.

The review of the literature shows how empirical scales of measurement were anticipated to represent great improvement over human intuitive

processes. It goes on to record how this challenge to human judgment was frustrated by the limitations inherent in the empirical approach itself.

Three pitfalls are well documented. The first is the finding that the empirical approach has created over-confidence in one's description of objects and events. The application of empirical scales to measurement brought about a feeling that one's description of the world is determined by nothing human but by something upon which subjectivity has no effect. And being so, the measurement made was thought to be perfect. This led many to be preoccupied by search for measurement scales which are more empirical and uncontaminated by man's subjective experiences. This view equates more empirical scales with more accuracy.

The extent to which the scale for measurement was thought to be free from human judgment itself became the criterion to evaluate empirical scales.

The second pitfall was that the empirical approach had to provide its own reference point for evaluating the adequacy of empirical scales and unwisely resort back to human judgment for this reference point. However, if human judgment is inadequate in the first place it cannot provide an objective reference on which to evaluate objective techniques.

It is no surprise therefore that finding independent criterion to judge between quantitative and qualitative approaches to measurement has been the greatest obstacle encountered in any endeavour to compare psychometric and professional techniques of testing and test constructions.

A third pitfall seen in the literature review concerns a general consensus of opinion that tests can be limited and abused, especially when administered and interpreted by unskilled personnel (Dyer, 1962).

This of course, tended to exclude the professional teacher from the assessment process.

On the other hand, the literature is replete with positive findings concerning objective measurement. There is evidence that tests are basically useful instruments and provide adequate information for decision-making situations (Dyer, 1962); that tests are preferable to other techniques of assessment such as subjective ratings, personal observations etc. Both standardized and teacher-made tests were seen to assess and provide relatively better information about student progress and learning difficulties. They also exposed the nature of the weaknesses of the learner, and the appropriate allocations of individuals into suitable programmes.

This research takes its starting point from the particular educational problems of assessment in Somalia, a system which is heavily dependent on teachers' judgments. The need to reform the system towards greater objectivity without a loss of any positive value currently present is the challenge presented to policy makers.

We are in a position to attempt an answer to the three questions which guided this research. From stage one, there is evidence that teachers see the objectives of assessment as wider than those normally encompassed by psychometric testing. Tests should take into account not only validity, reliability and objectivity but also personal circumstances and knowledge of the individual being tested. They should show recognition of the fact that objectivity is not equated with empirical evidence. It does not matter whether one employs an empirical technique or a human judgment. With either method one can be more or less objective. To be objective in this sense is to consider all relevant factors in preparing and constructing the test.

In the attitude questionnaire, it was argued that teachers' attitudes towards methods of assessment influenced the school examination system. Teachers' attitudes towards methods of testing can affect the construction, selection and administration of tests. It was also thought that teachers' positive or negative attitudes towards particular types of tests to some extent reflected the amount of experience they had had with those particular tests. Three important questions had to be answered in this part of the study. These questions were:

(a) Do teachers, in general, have more favourable attitudes towards psychometric tests than professional tests?

(b) Do teachers' favourable responses towards psychometric tests reflect knowledge of psychometric tests?

(c) Do teachers' responses towards psychometric tests differ according to certain background variables?

With regard to the first question, teachers were more positive towards psychometric tests, testing and test-construction procedures. The responses of teachers who claimed a higher degree of familiarity with the construction and administration of psychometric tests have been analysed and compared with the responses of teachers who claimed no familiarity. Teachers who have taken courses in educational measurement expressed more favourable responses towards psychometric tests.

Subjects did not differ in their responses according to sex, the subject taught, the number of years the teacher has been in the profession, the number of years the teacher has been teaching the subject, or the age of the pupils taught. These results were confirmed in a second and independent administration of the attitude scale. For teachers with more familiarity with educational measurements, the psychometric tests were

preferable to the professional tests. Goslin (1967) similarly concluded that teachers who expressed greater confidence in more objective tests were those with greater psychometric experience.

Apart from such factual details the data were analysed for deeper insight into the way teachers were formulating their judgment as indicated earlier. Thus the first stage of the study provided a list of criteria of judgment as well as a set of verifiable statements about teachers' reactions. It could be reasonably argued that the former have greater importance since they offer the possibility of explaining why teachers feel the way they do about these tests.

When constructing test items, the constructor (teacher) encounters many variables which influence the qualities of the item. He has, for example, to consider the content relevance in relation to what has been taught in class; the difficulty level of the item in relation to the abilities of the testees; and the discriminating power of the item in relation to the perceived range of abilities of the persons to be tested. Other variables to be accounted for include assumptions made about the testees' background, effectiveness of teaching, accurate knowledge of the abilities to be assessed and so on.

The question to be asked is can these be accomplished subjectively by judgment only. In a multivariate judgmental task of this kind it is possible that a teacher loses control and coordination of these multiple tasks and hence fails to account for some important variables.

Is teachers' ability to judge test item properties adequate? If teachers are able to judge test item properties, how well do they judge these properties? These are the questions to be answered in this part of the study.

To answer these questions teachers' judgments of psychometric properties of 24 test items have been recorded. Teachers' ratings of the item properties have been correlated with calculated empirical values obtained for the same item properties from student performances. The degree of relationship between teachers' ratings and the empirical psychometric properties of the test items is taken to indicate the adequacy of teachers' estimates of the item properties.

Of the 144 correlation coefficients obtained only 2 are significant (.05). The results indicate that teachers, as a group, were not able to judge test item properties adequately. The sizes and the patterns of the correlation coefficients further indicate that teachers' ratings of the test properties were not connected with the actual values obtained from student performances any better than would have been expected by chance.

Our conclusion is that teachers' judgment of the test item properties are not adequate.

A second question to be answered was which of the teachers' ratings of difficulty, discrimination, content relevance and item stability is more related in the overall assessment of the test item quality? For the psychometrician the test item property that weighs most is the item validity. Psychometricians argue that when the validity of the item is adequate other item properties are generally satisfactory. To determine the relative contributions of these item properties to the overall rating of the item, a stepwise multiple regression analysis has been used to discover which test item property predicts or explains more variation in the overall rating.

From these results it was concluded that teacher ratings of difficulty, discrimination, and stability depend very much on teachers'

perception of the item's content relevance to what has been taught in the class. Content relevance of the test item has more relationship with the overall assessment of the item. The relationship of other item properties with the overall item depend on the item's current relevance. In most of the cases these properties had a relationship only through the item's content relevance.

Teachers' judgments of item properties reflect neither the psychometric and objectively computed values nor the actual performance of their pupils. Further analysis of these judgments throw some light on the underlying judgmental process employed by the teachers. They regard item relevance as the major contributor to the process of assessment and related other properties to it. This is slightly at variance with the principles advocated by psychometricians who do not limit validity to item relevance only.

Having studied the teachers' attitudes towards tests and their construction and the teachers' ability and approach in judging the properties of test items, the study turns to the reactions of the consumers of tests, i.e. the pupils themselves. We asked do pupils show a preference for tests constructed by psychometricians or by teachers? Are there other criteria perceived by pupils as more important than the objective ones psychometricians call for? If so, could it be that in spite of teachers' demonstrated inability to judge the appropriateness of items for a test that pupils feel more comfortable and satisfied with teacher-constructed tests? To answer this question, the final stage of the investigation compared the relative worth of psychometric and professional tests as judged by students.

Within the limits of the sample and subject matter employed, it appears that teachers can generate tests equal in validity, fairness to

the examinees and reliability to psychometrically constructed tests. There are no significant differences between the reactions to the two types of test either on actual performance or on their ratings.

The difference in qualities between psychometrically constructed tests and the professionally constructed tests cited by the literature is probably due to the fact that they construct them under different conditions using different procedures. The research literature on the relative worth of the two types of test did not cite a single case where the professional test constructors and psychometric test constructors were put under the same constraints to generate test items.

One pitfall of the comparisons is that psychometric tests are evaluated on unidimensional measures. The literature fails to justify the use of unidimensionality as applied to educational achievement. Yet the comparisons are made judging teacher-made tests on the same criterion. On the other hand, the objectives of the teachers are not guided by unidimensionality. The psychometric test constructor is constrained to generate a unidimensional measure. Even when the psychometric test is intended to measure several characteristics, it is designed in such a way that each characteristic is measured by a sub-scale which is unidimensional. Therefore, once difference in quality between the two types of test lies in the difference of objectives rather than difference in ability to construct tests.

There is no evidence that costly statistical manipulation of items adds significantly to their validity, reliability, fairness or acceptability to the students. Apart from the need for replication of this finding, one still hesitates to recommend a total change to psychometric procedures in a country like Somalia with its more subjective traditions. One feels on firmer ground suggesting more training for would-be question setters and examiners rather than statistical manipulation.

APPENDIX I

Teachers' Attitudes towards Psychometric Versus Professional Procedures of Assessing a given Characteristic of a Class of Students

First Administration

Information About The Teacher

1. Sex: Male ____ Female ____

2. Number of Years in the profession: ____years ____months

3. Subject taught: Languages ____ Social Science ____ Maths ____
   Others (please state) _____

4. Number of years you have been teaching the present subject:
   ____ years ____ months

5. Number of years/months you have been teaching the present class of children:
   ____ years ____ months

6. The approximate size of the class: ____ students

7. Age range of the pupils in the class: ____ to ____

8. How many courses in each of the following areas have you studied?
   Intelligence testing ____ Achievement testing ____ Diagnostic testing ____ Personality testing ____ Others (please state) _____
   None _____

Our aim is to know when and what condition(s) do most teachers incline toward more professional/Psychometric procedures of assessing pupils in a class. What we mean by psychometric procedure of test construction is that process in which test items have been written, tried out on a sample of subjects and then subjected to statistical analysis so that all test item properties such as difficulty, discrimination, validity, reliability etc. for each item, become known in advance.

By profession, we mean a classroom teacher-made test or teacher's own personal ratings of the pupils which have not been subjected to statistical analysis so that test item properties are not known in advance.

Now, suppose you were asked to measure a given ability of characteristic of children in your class, say their arithmetic attainment. Assume naturally that you have knowledge of the children's age, educational level, linguistic ability, social background etc. Keeping that class of children in mind, would you please, answer the following statements by ticking the appropriate column: Strongly agree, agree, uncertain, disagree or strongly disagree. For example, someone in favour of the

first statement below but who feels that there are some exceptions would tick as follows. Similarly, someone who doesn't particularly like children could disagree with the second statement and tick the strongly disagree column.

Examples	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
Children bring a husband and wife closer to each other					
On balance, children are more of a blessing than a burden.					

The Attitude Scale

STATEMENTS	RESPONSE CATEGORIES				
	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
1. Results of standardized tests often substantially differ from the true ability of pupils as I observe them in the classroom					
2. Psychological tests sample the relevant content better than classroom teacher-made tests.					
3. I can predict my pupils' performance in the GCE examinations better by giving them a classroom test than by giving them standardized objective test					
4. Teachers should always bother building a table which matches course content with the instructional objectives before constructing tests					
5. It is unfair to use personal judgments as a measure of students' attainments since they are influenced by other characteristics of the pupils					
6. I believe that I know my students better than any test can tell					
7. Parents trust standardized tests more than they would trust teachers' judgments of their children as the basis for assigning course grades					
8. Given two scores of a student's academic ability I would rather assign more weight to the assessment of the psychologist than the professional assessment of the classroom teacher					

THE ATTITUDE SCALE

STATEMENTS	RESPONSE CATEGORIES				
	Strongly agree	Agree	Uncertain	Disagree	Strongly disagree
9. Matching psychologically designed test items with the subject matter taught by the teacher is equally difficult as constructing a new test					
10. On the whole, psychological tests are not better than personal judgments because all the subjectivity originally avoided creeps in in the process of selecting appropriate items from standardized tests, administering and scoring.					
11. Students trust standardized tests more than they would trust teachers' judgments of their attainments as the basis for assigning course grades					
12. Tests constructed by classroom teachers are not as good as tests constructed by school psychologists					
13. It is not fair to compare the academic attainments of pupils from different classes or regions on the basis of measures (grades) obtained by teachers' own personal judgments of the pupils					
14. Psychological tests to not measure the higher mental processes such as understanding, interpretation, application etc. but only the factual information such as rote recall of facts, dates, places etc.					
15.Standardized tests free teachers or other professionals for other more scholarly and creative duties in educating students.					

STATEMENTS	RESPONSE CATEGORIES				
	Strongly agree	Agree	Uncertain	Disagree	Strongly disagree
16. I believe all test items should be subjected to statistical analysis before I administer them to a class.					
17. Each school should have its psychologist resonsible for only testing and test construction.					
18. I trust my own personal ratings of my class less than I would trust a test constructed by a psychologist.					
19. It is fair to compare the academic attainments of pupils from different classes, or regions on the basis of psychologically constructed tests.					
20. One should trust in standardized tests less than teacher-made tests or teachers' own personal judgments of the pupils because teachers can prepare their students specially for these standardized tests.					
21. Ambiguity is a common fault of many classroom teacher-made tests when compared with the tests constructed by psychologists.					
22. Psychologically constructed tests are less reliable and less efficient as selective instruments than interviews and teachers' recommendations.					

THE ATTITUDE SCALE

STATEMENTS	RESPONSE CATEGORIES				
	Strongly agree	Agree	Uncertain	Disagree	Strongly disagree
23. Since their test items will be anyway subjected to rigorous item analysis later on, psychologists should not worry about the initial construction of the test items.					
24. Giving students practice on standardized tests is not the responsibility of teachers.					
25. Without using any test, I could rank nearly all my class according to their knowledge of the subject I teach					
26. Compared to psychological tests, school grades and teachers' estimates of academic attainments are more influenced by pupils' social and cultural background					
27. Psychological tests are less valid and reliable than classroom tests or judgments.					
28. Teacher-made tests do not cover the objectives taught by the teacher neither do they reflect the objectives proportionally					
29. No valid test can be constructed without following a table of test specification.					

THE ATTITUDE SCALE

STATEMENTS	RESPONSE CATEGORIES				
	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
30. Teachers preparing their students for standardized tests negatively interferes with external examinations.					
31. I believe I could assign overall grades to my pupils without giving them a written test					
32. Standardized objective tests are indifferent to the pupils' social, cultural, regional and racial back-grounds.					
33. I believe there is no harm in assigning course grades partially on the basis of earlier standardized test scores of the students.					
34. A great difficulty in teacher-made tests, compared with standardized tests is that there is no way of determining in advance whether the test is too difficult or too easy for the students					
35. Teachers evaluation tools such as teacher-made tests, rating scales, anecdotal records, observational techniques etc, are better predictors of student progress					
36. Teachers can construct better tests for their own classrooms than any test technician can do.					
37. Pupil evaluation should not be left to the test psychologist alone, but teachers should be equally active part-icipants in the process of selecting tests for the school.					

THE ATTITUDE SCALE

STATEMENTS	RESPONSE CATEGORIES				
	Strongly agree	Agree	Uncertain	Disagree	Strongly disagree
38. I feel that an experience of two or more years with the class would allow the teacher an accurate estimate of his students' academic attainments					
39. To do justice to all students, all tests should be constructed statistically by a test psychologist.					
40. I think the single most accurate measure (test) of a student's intellectual and academic ability is the psychologically constructed test of the psychologist					
41. Standardized objective tests are less economical, in terms of time and money, than the subjective measures of assessment.					
42. Statistical refinement of test items do not ensure the usefulness of a test					
43. Communication among the professionals becomes more precise and understandable when standardized tests are used than when personal evaluations are used.					
44. We must always substantiate subjective judgments in teacher-made tests by a statistical analysis of the test items.					

THE ATTITUDE SCALE

STATEMENTS	RESPONSE CATEGORIES				
	Strongly agree	Agree	Uncertain	Disagree	Strongly disagree
45. In psychological tests, since candidates do attempt the same questions, we can rightfully compare the performance of the pupils in the same class.					
46. Since teachers know the relative abilities of all the children in their classes, written tests are only to satisfy parents, or the public, about the fairness of the grades given.					
47. Unless compelled otherwise I will use essay type questions most of the time					
48. Test item writing is an art that requires special abilities of the test constructor					
49. In an essay type test, since candidates do not attempt exactly the same questions, we cannot compare the performance of the pupils in the same class					
50. The teacher, however long he was teaching the class, can only know the abilities of few pupils in the class while the abilities of the majority are blurred.					
51. It is fair to rank order a class of students entirely by using a standardized objective test.					

THE ATTITUDE SCALE

STATEMENTS	RESPONSE CATEGORIES				
	Strongly agree	Agree	Uncertain	Disagree	Strongly disagree
52. School psychologists have no vivid vision of subjects in their minds when they are constructing the test, as teachers do.					
53. The process of obtaining standardized tests, understanding the instructions of administering to subjects, and scoring them is more tedious than constructing teacher's own test.					
54. The kinds of abilities measured by standardized tests are not important in determining subsequent academic success of children.					

APPENDIX II

SCALE FOR TEACHERS' JUDGMENTS OF TEST PROPERTIES

1. To which of the following food groups does starch belong?

   A. Protein
   B. Carbohydrate
   C. Fat
   D. Vitamin
   E. Mineral Salt

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

2. In mammals energy is released from

A. Oxygen
B. Water
C. Haemoglobin
D. Glucose
E. Carbon dioxide

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

3. Wind pollinated flowers differ from insect pollinated in that they

A. are brightly coloured
B. possess scent
C. produce larger quantities of pollen
D. have shorter filaments
E. produce larger quantities of nectar

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

4. The fertilized ovum is called

A. a zygote
B. an ovule
C. a placenta
D. a sperm
E. an ovary

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

5. Bones and muscles are attached to each other by

A. cartilage
B. joints
C. ligaments
D. nerves
E. tendons

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

6. The part of the ear concerned with balance is the

A. ear drum
B. ear ossicle
C. semi-circular canal
D. cochlea
E. eustachean tube

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

7. Bile is

A. a digestive enzyme
B. an emulsifier of fat
C. a chemical found in throat
D. a chemical which breaks down carbohydrate
E. an important food substance

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

8. Which of the following is not essential for animals?

A. Hydrogen gas (H_2)
B. Oxygen gas (02)
C. Iron (Fe)
D. Magnesium (Mg)
E. Calcium (Ca)

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

9. Which of the following does the environment Not provide for a growing plants.

A. Light

B. Oxygen

C. Nitrates

D. Water

E. Carbohydrates

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

10. Of the following acids, the one that helps in the process of digestion of food is

A. lactic acid

B. hydrochloric acid

C. amino acid

D. acetic acid

E. fatty acid

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

11. Which of the following vitamins can be formed in the human skin?

A. Vitamin A
B. Vitamin B complex
C. Vitamin C
D. Vitamin D
E. Vitamin K

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

12. Which of the following plants growing on a lawn is a monocotyledon?

A. moss
B. grass
C. dandelion
D. daisy
E. clover

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

13. When soil is shaken up in water, the part which stays in suspension in the water is the

A. humus

B. mineral salts

C. sand particles

D. clay

E. grit particles

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

14. Which of the following parts of a cell is usually only found in a plant cell?

A. nucleus
B. cell membrane
C. cytoplasm
D. large vacuole
E. chromosome

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

15. Shortage of iodine in the diet may cause the body to make too little

A. thyroxine
B. insulin
C. adrenalin
D. urine
E. bile

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

16. Of the following processes, the one that should kill all bacterial life present is

A. pasteurisation
B. autoclaving with a pressure cooker
C. deep freezing
D. boiling for five minutes
E. washing with disinfectant

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

17. Which of the following foods which we obtain from plants first requires the plant to be pollinated and fertilised?

A. cabbage

B. garden peas

C. celery

D. cauliflower

E. carrot

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

18. Which of the following qualities is determined only by the chromosomes?

A. Body height

B. personality

C. body weight

D. eye colour

E. intelligence

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

19. After allowing for different body sizes which of the following animals will lose most water through the kidneys?

A. gerbil

B. starling

C. mouse

D. herring

E. goldfish

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

20. More fish species now live in the River Thames as it flows through London than were found ten years ago. This is chiefly because

A. fish have become used to living in dirty water

B. less shipping means less disturbance of the water

C. the water contains more dissolved oxygen now

D. the water is less dirty than ten years ago

E. less rubbish is thrown into the water now

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

21. When a person touches a hot object, which of the following parts of the nervous system is NOT directly involved in the reflex response?

A. sensory neurons
B. receptor organ
C. brain
D. connector neurons
E. Motor neurons

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

22. Which of the following is NOT a part of a terminal bud such as a horse chestnut bud?

A. growing point

B. flower buds

C. fruits

D. tiny foilage leaves

E. scales leaves

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

23. Which of the following would a biologist say is NOT a true fruit?

A. cherry
B. tomato
C. cucumber
D. melon
E. rhubarb

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor	Poor	Fair	Good	Very good

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same	Slightly better	Somewhat better	Much better	Very much better

D. How relevant is this examination question to the material you taught in class?

Not at all relevant	Somewhat relevant	Relevant	Quite Relevant	Very much Relevant

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

24. Which of the following does NOT apply to lymph?

A. bathes the body cells
B. carries dissolved oxygen
C. contains red corpuscles
D. contains white cells
E. flows into blood system

A. How good is this examination question for assessing your pupils' understanding of the material you taught?

Very poor		Poor		Fair		Good		Very good	

B. What percentage of the pupils in the class will answer this question correctly?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. Think of your class divided according to ability into a top, middle and a bottom stream. How much better will the top stream do on this examination question compared to the bottom stream?

Same		Slightly better		Somewhat better		Much better		Very much better	

D. How relevant is this examination question to the material you taught in class?

Not at all relevant		Somewhat relevant		Relevant		Quite Relevant		Very much Relevant	

E. If your class had to answer the same question on two occasions, 7 days apart, what percentage of those who passed on the first occasion would pass again on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

F. What percentage of those who failed on the first occasion would pass on the second occasion?

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

APPENDIX III

INSTRUCTIONS TO ITEM WRITERS

"We intend to develop tests for pupils taking the CSE or the GCE Examinations. To create sufficient item banks in biology we need more items for pre-testing. You are required to contribute to the item bank by drafting a number of multiple choice biology questions. The distribution of the items across ability categories and the syllabus sections should be proportional to that prescribed by the syllabus. In the form attached, please enter the following information: Write the ability measured by the test item and the section of the syllabus it represents. Indicate the proportion of students which you think will pass the item, or rate each item as easy, average or difficult as it will appear to an average O-Level candidate. On the whole, the test items you will draft should apparently have an average difficulty level. Please indicate the correct key (A,B,C,D,E) to each item by entering the letter which corresponds to the correct answer in the form. ALL options should be equally attractive to the candidates."

Teachers' Booklet for Multiple-Choice Test

Additional material provided to panel members

was 'Teacher's Booklet For the Multiple-Choice Objective Test (Biology)'. We have also assumed that most of our panel members were already familiar with the booklet. Since 1976, the University of London GCE O-Level Biology paper I has been a multiple choice test. One purpose of the booklet was to explain to biology teachers' reasons for the introduction of multiple choice test. A second purpose of the booklet has been to give candidates for the examination and the teachers an idea about the types of question which are included in the examination. Other reasons for introducing the multiple choice test are outlined in

the booklet as follows:

"1. The test (the multiple choice test) is constructed according to an agreed specification and a definite weighting can be given to each syllabus section and ability. This allows a year to year consistency to be maintained.

2. All the questions in London GCE multiple choice will have been pre-tested and therefore, there is reasonable certainty that the questions are free from ambiguity, are of the appropriate level of difficulty and will discriminate well between the candidates.

3. Multiple choice questions do not require written answers and this allows a good coverage of the syllabus to be examined in a relatively short time. Candidates need to expend little or no time in writing so may devote most of their time to thinking.

4. The mark gained by a candidate is simply the total number of questions he has answered correctly. The marking is, therefore, objective and entirely free from subjective judgment.

5. Multiple question tests, as with all other types of compulsory short answer tests, make the examinations fairer and more reliable since in each test all candidates answer the same questions."

Most of these reasons have been mentioned earlier.

THE PROCEDURE AND THE DEVELOPMENT OF PSYCHOMETRIC TESTS

1. First, the would-be item writers are given some training. Once trained, the teachers are assigned to panels of item writers who are commissioned to draft multiple choice test items.

2. A second group reviews and edits the panel drafted multiple choice test items. This group is composed of a moderator, chief examiners, and teachers. According to their personal judgments, the group accepts some of the test items and rejects others.

3. Acceptable items are pre-tested on a representative sample of subjects. Scores of the candidates on the test items are analysed and the relevant test item properties are computed for each item. Only those items which satisfy certain criteria are included in the final ensemble of a psychometric test item. These criteria for the item selection are found in books on educational and psychological measurement. Our interest in the above procedure ends at this point.

TEST SPECIFICATION

Some other information contained in the booklet was a sample of test specifications. We have already mentioned the importance of the instructional objectives to test constructors and to others evaluators of educational achievements. The test specification given in the booklet was a two-way table of abilities (as classified by Bloom, 1956) by syllabus (content) on which candidates are to be tested. In a sense, these abilities stood for the educational objective of the GCE. Desired proportions of test items were assigned to each ability by syllabus section.

EXAMPLES OF MULTIPLE CHOICE QUESTIONS

"Directions:

Each of these questions or incomplete statements is followed by five suggested answers. Select the best answer in each case and mark the sheet appropriately."

Example question 1:

In a parasitic relationship between two organisms

A both members benefit

B neither is harmed

C both members suffer harm

D only one member gains benefit

E neither member gains benefit

Statistical analysis of question 1

Students classified by total test score	A	B	C	D	E	Omits
lowest fifth	5	3	3	40	4	0
next lowest fifth	1	5	1	48	0	0
middle lowest fifth	1	1	1	51	1	0
next highest fifth	1	0	0	53	0	0
highest	1	3	0	52	0	0
TOTAL	9	12	5	244	6	0

* correct answer

facility(diff.) = 0.88
discrimination = 0.39

Syllabus section = 3b
Ability = comprehension

Comment:

'A very easy question for this group of candidates with 88% answering correctly. The discrimination was satisfactory. The ability tested was considered to be simple comprehension rather than just knowledge because of the wording of the options which requires some degree of comparison'.

Example question 3:

A farmer intends to devote a hundred acres of grassland to production of food in the form of protein. Which of the following factors would have least effect on his productivity?

A amount of light

B temperature

C type of animal reared

D type of soil

E amount of rainfall

Statistical analysis of question 3

Student classified by total test score	A	B	C*	D	E	Omits
lowest fifth	8	5	37	2	2	1
next lowest fifth	9	5	33	5	3	0
middle fifth	8	7	35	3	2	0
next highest fifth	6	8	35	4	2	0
highest fifth	7	14	26	8	1	0
TOTAL	38	39	166	22	10	1

Facility = 0.60
Discrimination = 0.15

Syllabus section = 3c
Ability = application

Comment: This was an unsatisfactory question. The pre-test statistics reveal a low level of discrimination. The mental processes involved in reaching the best answer were complicated by the negative nature of the question asked. This is an unsatisfactory question as it is difficult to arrive at the right answer.

Example question 12:

Which of the following is found on woody stems:

A axil

B lenticel

C midrib

D node

E petiole

Statistical analysis of question 12

Students classified by total test score	A	B	C	D	E	Omits
lowest fifth	6	13	10	13	12	2
next lowest fifth	4	22	4	11	13	3
middle fifth	9	21	6	8	10	1
next highest fifth	6	34	1	0	7	0
highest fifth	4	49	1	8	1	0
TOTAL	29	139	27	40	43	6

facility = 0.49
discrimination = 0.55

Syllabus section
Ability = knowledge

Comment: The pre-test statistics on question 12 are very satisfactory. Half of the candidates chose the correct answers and these tended to be the better candidates as judged by their showing on the whole pre-test paper. All the incorrect options proved to be suitable distractors.

ADDITIONAL INFORMATION GIVEN TO TEACHERS

Ability	Definition
"Knowledge	The ability to recall facts, nomenclature, classification, practical techniques etc.
Comprehension	The ability to calculate, to translate data from one form to another (verbal into mathematical or graphical) to interpret and deduce the significance of data and to solve problems in which the mode of solution of the problems should be familiar.
Application	The ability to apply knowledge, experience and skill to new situations presented in a novel manner.
Analysis/Evaluation	The ability to analyse given information into its various parts and, as a result, to make a judgment as to its value."

The definitions were given to ensure better comparability of all 4 tests. According to Bloom (1956) items which measure different abilities

have usually different levels of difficulties. Different levels of difficulties pose problems of comparability of tests (Gullikson, 1965). To make the number of test items measuring each ability and the number of items representing each section of the syllabus similar, item writers were specifically instructed to produce the desired distribution of test items over ability categories and over sections of the syllabus. All the tests obtained had the desired distributions of items over ability categories and over sections of the syllabus. Hence, test items used were balanced in terms of their subject-matter content and in terms of the behavioural content.

BIBLIOGRAPHY

ADKINS, D.S. Test Construction. Second edition. Charles E. Merrill Pub. Co. Colombus Ohio: 1974.

ANASTASI, A. Psychological Testing (3rd) New York: Macmillan, 1968.

ANASTASI, A. Psychological Testing (4ed) New York: Macmillan, 1976.

ANGOFF, W.H. (1971) 'Scares, norms and Equivalent scores' In Educational Measurement. 2nd ed. Washington, D.C.: American Council on Education, 1971.

ANNETT, J. Psychometric. Open University, 1974.

AUSUBEL, D.P. & ROBINSON, F.G. School Learning. New York: Holt, Rinehart and Winston, 1969.

BARCLAY, J. Controversial Issue in Testing. New York: Houghton Mifftin Co., 1968.

BERK, R.A. 'Item Analysis'. In Criterion-Referenced Measurement. R.A. Berk (ed.) Johns Hopkins Univ. Press, 1980.

BERK, R.A. Determination of optional cutting scores in criterion-referenced measurement. Journal of Educational Measurement, 1976.

BLOOD, D.I. & BUDD, W.C. Educational Measurement and Evaluation. New York: Harper and Row, Pub. 1972.

BLOOM, B. Taxonomy of Educational Objectives. London: Longman, 1956.

BRENNAN, R.L. & KANE, M.T. (1977a) 'An Index of dependability for mastery tests' Journal of Educational Measurement. 14, 277-287.

BRENNAN, R.L. & KANE, M.T. (1977b) Signal noise ratios for domain-referenced tests. Psychometrika, 42, 609-25.

BRENNAN, R.L. (1977a) Generability Analysis: Principles and Procedures. Act Technical Bulletin No. 26 Iowa City, Iowa: American College

BRENNAN, R.L. (1977b) KR-21 and lower limits of an index of dependability for mastery tests. Act Technical Bulletin No. 27, Iowa City, Iowa: American College Testing.

BRENNAN, R.L. (1978) Extensions of Generalizability theory to domain-referenced testing. Act Technical Bulletin. No. 30, Iowa City, Iowa: American College Testing Program.

BRENNAN, R.L. (1980) 'Applications of generalizabilit y theory'. In. Criterion-Referenced Measurement, R.A. Berk, ed. 1980. Johns Hopkins University Press, Baltimore, 1980.

BROADFOOT, P. (1979) Assessment, Schools and Society, London: Methuen.

BROWN, F. (1970) Principles of Educational and Psychological Testing. Hinsdale, Illinois: Dryden Press, 1970.

BURROUGHS, G.E.R. (1975). Design and Analysis in Education Research. Oxford: Alden and Mowbray.

BURNS, R.B. & DOBSON, C.B. (1981) Experimental Psychology: Research Methods and Statistics. Lancaster: MTP Press.

CARVER, R.P. (1970). Special Problems in measuring change with psychometric devices. In Evaluative Research Strategies and Methods. Pittsburgh, Pa.: American Institute for Research, 48-63.

CAMPBELL, J. & JANSTED, K. 'The effect of group participation on brain storming effectiveness for two samples'.

CAMPBELL, N.R. (1938) Symposium: Measurement and its importance for philosophy. Aristotelian Society, Suppl. Vol. 17, London: Harrison.

CHILD, D. Psychology and the Teacher. London: Holt, Rinehart and Winston, 1973.

COFFMANN, W.E. (1971a)'Essay examinations' In Educational Measurement. R. Thorndike (ed.) Washington, D.C. American Council on Education.

COFFMAN, W.E. (1971b) On the Reliability of Rating Essay Examinations. Research in the Teaching of English.

CRONBACH, L. (1970) Essentials of Testing. Harper-Row.

CRONBACH, L.J. (1971) Test validation. In Educational Measurement. R.Thorndike (ed.) 2nd ed. Washington D.C.: American Council on Education.

CRONBACH, L.J., Glaser, G.C., NANDA, H. & RAJARTNAM, N. (1972) The dependability of behavioural measurement: Theory of Generability for Scores and Profiles,

CUREON, E.E. (1957) The Upper and Lower Twenty Seven Percent Rule. Psychometrika, 22, 293-296.

DU BOIS, P.H. A Test Dominated Society: China, 115BC-1905 A.D. In: Anne Anastasi, (ed.) Testing Problems Perspective. Washington D.C.: American Council on Education, 1964.

DUNNETE, (1963). Journal of Applied Psychology 47 (1), 30-37.

EBEL, R.L.(1962a) Content Standard Scores. Educational and Psychological Measurement. Vol. XXII, No. 1, 1962.

EBEL, R.L. (1962b) Measurement and the Teacher. In Educational and Psychological Measurement. Morriston,New Jersey, General Press.

EBEL, R.L. (1971) Criterion-Referenced Measurement Limitations. School Review, 1971, 69, 282-288.

EBEL, R.L. (1975) The social consequence of Educational Testing. In Educational and Psychological Measurement. Payne, D. & McMorris, R. (ed.) Morristown: General Learning Press.

EBEL, R.L. (1979) Essentials of Educational Measurement. 3rd ed. Englewood Cliffs, N.J.: Prentice-Hall.

ELLIS, B.D. (1966) Basic concepts of measurement. Cambridge: Cambridge University Press.

FERGUSON, G.A. (1976) Statistical Analysis in Psychology and Education. Tokyo: McGraw-Hill.

FITCH, M.L., DRUCKER, A.J., NORTON, J.A. (1951) Frequent testing as motivating factor in large lecture classes. Journal of Educational Psychology, 42, 1-19.

FITZPATRICK & MORRIS (1971), 'Performance and production evaluation'. Educational Measurement, American Council on Education, D.C.: Washington 1971
By Thorndike, R. L.

FRASER, D.O. Measurement in Psychology. The British Psychological Society, 1980.

GLASER, R. (1962) Content Standard Test Scores. Educational and Psychological Measurement, 22, 15-25.

GLASER, R. (1963). Instructional technology and the measurement: some questions. American Psychologist, 18.

GLASER, R. & NITKO, A.J. (1971) Measurement in learning and instruction. In Educational Measurement. R.L. Thorndike (ed.), Washington: American Council on Education.

GLASS, G.V. (1978) Standards and Criteria. Journal of Educational Measurement. Vol. 15, no. 4.

GOSLIN, D. Standardized Ability Tests and Testing. In Educational and Psychological Measurement. Payne, D. & McMorris (ed.) Morristown: General Press, 1975.

GRONLUND, N.E. Measurement and Evaluation in Teaching. New York: Macmillan, 1965.

GRONLUND, N.E. (1970). Measurement and Evaluation in Teaching. New York: Macmillan.

GRONLUND, N.E. (1971) Measurement and Evaluation in Teaching. New York: Macmillan.

GRONLUND, N.E. (1981) Measurement and Evaluation in Teaching. New York: Macmillan.

GUILFORD, J.P. Psychometric Methods: New York: McGraw-Hill, 1965.

GUILDFORD, J.P. & FRUCHTER, B. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill, 1978.

GUIEN, R.M. (1977) Content Validity; the source of my discontent. Applied Psychological Measurement. 1:1-10.

GULLIKSON, H. (1965) Theory of Mental Tests. New York: John Wiley.

HAMBLETON, R.K. (1978) On the use of cut-off scores with criterion-referenced tests in instructional setting. Journal of Educational Measurement. 15: 277-290.

HAMBLETON, R.K. (1980) Test score validity and standard-setting methods. In Criterion-Referenced Measurement. The Johns Hopkins University Press, Baltimore.

HAMBLETON, R.K. & NOVICK, M.R. (1973) Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement. 10: 159-170.

HAMBLETON, R.K., SWAMINATHAN, H., ALGINA, J. & COULSON, D.B. (1978). Criterion-referenced testing and measurement. A review of technical issues and development. Review of Educational Research 48:1-47.

HELMSTADTER, G.C. Principles of Psychological Measurement. New York: Appleton-century-crofts, 1964.

HENRYSON, S. (1971) Gathering, analysing and using data on test items. In R.L. Thorndike (ed.) Educational Measurement. Washington D.C.: American Council on Education.

HIVELY, W., PATTERSON, H.L. & PAGE, S.A. (1968) Universe-defined system of arithmetic achievement tests. Journal of Educational Measurement. 5:275-290.

HOFFMAN, B. (1962) Tyranny of Testing. New York: Crowell-Collier, Macmillan.

HOLDER, O. (1901). Die Axiome de quantitat die lehre von mass. Berichte ueber die verhandluger der konogdich sachsischen gessell shaft dae wissenschaften zu leigzis, Mathematisch-Physische Class, 53:1-64.

HORROCKS, J.E. & SCHOONOVER, T.I. (1968) Measurement for Teachers. Colombus, Ohio: Charles E. Merril Pub. Co.

HUDELSON, E. The effect of objective standards upon composition teacher' judgments. Journal of Educational Research. Vol. 12 1925 (329-340).

HULL, C.H. & NIE, N.H. (1981) SPSS, Update 7-9. New York: McGraw-Hill.

HULTE[illegible], C.E. (1925) Personal elements in teachers' marks. Journal of Educational Research vol. 12.

HUYNH, H. (1976) On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement. 13:265-276.

JENSEN, A.R. (1980) Bias in Mental Testing. London: Methuen.

JONES, H. (1923) Experimental Studies of College Teaching. Archives of Psychology, Vol. LXVIII, November, 1923, pp. 36-70.

KELLEY, T.L. (1939) The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 30, 17-24.

KERLINGER, F.N. Foundations of Behavioural Research. New York: Holt-Sanders International, 1973.

KIRKPATRICK, J.E. The motivating evfect of a specific type of testing programme. University of Iowa. Studies in Education. Vol. IX, 1934, 41-69.

KRATZ,D.H. et al. (1971) Foundations of Measurement. Vol. 1. New York: Academic Press.

LENNON, R.T. (1956) Assumptions Underlying the Use of Content Validity. Educational and Psychological Measurement, 16:294-304.

LINDVALL, D.M. & NITKO, A.J. (1969) Criterion-Referenced Tests. Los Angeles: National Council on Measurement in Education.

LINN, R.L. (1979) Issues of validity. In Measurement for comperence based programs. In Practices and Problems in comperence-based Measurement. M.A. Bunda & J.R. Sanders (ed.) Washington D.C.: American Council on Education.

LORGE, I. & KRUGLOVE, L. (1953) The improvement of estimates of test difficulty. Educational and Psychological Means. Vol. 13, 34-46.

LUEPTOW, L.B., EARLY, K. & GARLAND, T.N. (1976) The validity of Student Evaluations of Objective Test Items. Educational and Psychological Measurement. Vol. 36, 939-944.

MAGER, R.F. (1962) Preparing Instructional Objectives. Palo Alto, CA: Feardon Pub.

MARSHALL, J.C. (1967) Composition errors and essay examination grades re-examined. American Educational Research Journal.4, 375-386.

MARSHALL, J.L. & HAERTEL, E.H. (1976) The mean split-half coefficient of agreement; A single administration index of reliability for mastery tests. Manuscript, University of Wisconsin.

MARTUZA, V.R. (1977) Applying Norm-Referenced and Criterion-Referenced Measurement in Education. Boston: Allyn and Bacon, Inc.

MASCUILO, L.A. & STAUGHTER, R.E. Statistical procedures for item bias. Journal of Educational Measurement. Vol. 18(4) 229-248.

MEHRENS, W. & LEHMANN, P. (1975) Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston

MEHRENS, W. & LEHMANN, P. (1978) Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston.

MESSICK, S.A. (1975) The Standard Problem Meaning and Values in measurements and Evaluation. American Psychologist, Vol. 30, 955-976.

MILLMAN, J. (1980) Computer-based item generation. In Criterion Measurement. R.A. Berk (ed.) Baltimore, Johns Hopkins University Press.

MOSTELLER, F. & TUKEY, J.W. (1977) Data Analysis and Regression. Cambridge, Massachussetts: Addison-Wesely.

NEDELSKY, L. (1954) Absolute grading standards for objective test. Educational and psychological measurement. 14:3-19.

NOLL, V.H. (1938) The effect of written tests upon achievement in college classes: an experiment and summary of Evidence. Psy. Bull. 938, 35, 671.

NUNNALLY, J.C. (1970) Introduction to Psychological Measurement. New York: McGraw-Hill.

NUNNALLY, J.C. (1972) Educational Measurement and Evaluation. New York: McGraw-Hill, 1972.

NUNNALLY, J.C. (1976) Psychometric Theory. New York: Macmillan.

NUNNALLY, J.C. (1978) Psychometric Theory. New York: McGraw-Hill.

OPPENHEIM, A.N. Questionnaire Design and Attitude Measurement. London: Heinemann, 1966.

PAYNE, D.A. & McMORRIS, R.F. Educational and Psychological Measurement. N.J. Morriston: General Learning Press, 1975.

POPHAM, W.J. (1978a) Criterion-Referenced Measurement. Englewood Cliffs N.J. Prentice Hall.

POPHAM, W.J. (1978b) Domain specification strategies. In Criterion-Referenced Measurement. R.A. Berk (ed.) Johns Hopkins University Press, Baltimore.

POPHAM, W.J. (1980) Domain specification Strategies. In Criterion-Referenced Measurement. R.A. Berk (ed.) Baltimore. Johns Hopkins University Press,

POPHAM, J. & HUSEK, (1969), Implications of criterion referenced measurement. Journal of Educational Measurement. Vol. 6, 1, 1-9.

PRYTULAK, L.S. (1975) Critique of S.S. Stevens' Theory of measurement scale classification. Perceptual and Motor Skills, 41, 3-28.

ROMIG, C.W. (1970) Attitudes of Classroom Teachers toward the roles of the school psychologist and other personnel of a psychological services centre for children. Dissertation Abstracts International. Vol. 31 (6-A), 2747.

ROSS, C.C. & HENRY, L.K. (1939) The relation between frequency of testing and progress in learning psychology. Journal of Educational Psychology, 30:603-611.

RUSSELL, B. (1938) Principles of Mathematics. 2nd. ed. New York: Norton.

RYAN, D.C., (1951) Results of internal consistency and external validation procedures applied in the analysis of test items measuring Professional Information. Educational and Psychological Measurement. 11, 549-60.

RYAN, J.J. (1968) Teacher Judgment of Test Item Properties. Journal of Educational Measurement. Vol. 5 No. 5 (301-306.

SNEDCOR, G. & COCHRAN, W. (1977) Statistical Methods. 6th ed. Ames, Iowa: Iowa State University Press.

STANLEY, J. Reliability. In Thorndike, R.L. (ed.) Educational Measurement. Washington D.C.: American Council on Education, 1971.

STANDLEE, L.J. & POPHAM, W.J. (1960). Psychological Rep., 6, 468.

STALNAKER (1951) The essay type examination in E.F. Lindquist, (ed.) Educational Measurement. Washington D.C.: American Council on Education, PO. 498-530.

STEVENS, S.S. (1951) Mathematics, measurement and psychophysics. In S.S. Stevens (ed.), Handbook of Experimental Psychology. 1-49, New York: Wiley.

STEVENS, S.S. (1958) Problems and methods of psychophysics. Psychological Bulletin, 55:177-196.

SUBKOVIAK, M.J. (1976) Estimating Reliability from a single administration of a mastery test. Journal of Educational Measurement. 13:265-276.

SUBKOVIAK, M.J. (1977) Veluation of Criterion-Referenced Reliability Coefficients. Final Report, Grant NIE-G-76-0088. Washington D.C. National Institute of Education.

SUBKOVIAK, M.J. (1978) Empirical Investigation of Procedures for estimating reliability for mastery tests. Journal of Educational Measurement.

SUBKOVIAK, M.J. (1980) Decision-consistency approaches. In Criterion-Referenced Measurement. R. Berk(ed.) The Johns Hopkins University Press, Baltimore, 1980.

SWAMINATHAN, H., HAMBLETON, R.K. & ALGINA. (1974) Reliability of criterion-referenced tests. A decision-theiretic formation. Journal of Educational Measurement. 11:263-67.

SYMONDS, P.M. (1928) Factors Influencing Test Reliability. Journal of Educational Psychology, Vol. 19: 73-87.

THORNDIKE, R.L. Reliability. In E.F. Lindquist (ed.), Educational Measurement. American Council on Education, 1951. Washington, D.C.

THORNDIKE, R.L. & HAGEN, E. Measurement and Evaluation in Psychology and Education. New York: John Wiley, 1955.

THORNDIKE, R.L. & HAGEN, E. Measurement and Evaluation in Education and Psychology, New York: John Wiley. 1977.

TORGERSON, W.L. (1958) Theory and Methods of Scaling. New York: Wiley.

TRAVERS, R.M. (1955) Educational Measurement. New York: The Macmillan Co.

TURNEY, A.H. (1931) The effect of frequent short objective tests upon the achievement of college students in educational psychology. School and society. Vol. XXXIII, 760-762.

WESMAN, A.G. (1952) Reliability and confidence. Test Service Bulletin, No. 44: May 1952.

WESMAN, A.G. (1976) Reliability and Confidence. In Readings in Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston,.

WESMAN, A.G. (1971) Writing the Test Item. In Thorndike, R.L. (ed.) Educational Measurement. Washington D.C.: American Council on Education.

WILMOT, J. (1975) Objective Test Analysis: some criteria for item selection. Research in Education. No. 13.

WOMER, F. (1975) Tests: Misconception, Misuse and overuse. In Educational and Psychological Measurement. Payne, D. & McMorris, (eds.). Morriston, New Jersey, General Press.

WOOD, D.A. (1962) Test construction: development and interpretation of Achievement Tests. Columbus, Ohio: Merrill.

ADDENDA

DYER, H. (1976) The menace of testing reconsidered. In: Readings in Measurement by W. Mehrens New York: Holt, Rinehart & Winston.

FLANAGAN, J.C. (1939) General considerations in the selection of test items. Journal of Educational Psychology. Vol. 30, (674-680).

POPHAM, W.J. (1981) Modern Educational Measurement. Englewood Cliffs: Prentice-Hall, L.

PYRCZAK, F. (1973) Validity of discrimination Index as a measure of item quality. Journal of Educational Measurement. Vol. 10, No. 3.

SEDDON, H. (1982) Poor Forecasting. Times Educational Supplement. (Feb)

ALLPORT, G.W. (1935) 'Attitudes' In C. Murchinson (ed.), A Handbook of Social Psychology. Worcester: Clark University Press

ANASTASI, A. (1965) (ed.) Individual Differences New York: John Wiley.

CATTELL, R. The scientific use of factor analysis. New York: Plenum, 1978.

CHILD, D. (1970) Essentials of Factor Analysis. London: Holt, Rinehart and Winston.

CONRAD, H.S. Characteristics and uses of Item Analysis data. Psychological Monographs, 1948, 62 (8, Whole No. 295).

CRONBACH, L. & MEEHL, P.E. Construct Validity in Psychological Tests. Psychological Bulletin, 1955, 52, 281-302.

EBEL, K.L. (1956) Obtaining and Reporting Evidence on Content Validity. Educational and Psychological Measurement, 269-282.

EDWARDS, A.L. (1957) Techniques of attitude scale construction. New York: Appleton-Crofts-Century.

GOSLIN, D. (1967). Teachers and Testing. New York: Russell Sage.

HAMBLETON, R.K. 'Competency Test Development, Validation and Standard Setting'. In Minimum Competency Achievement Testing. (ed.) R.M. Jaeger and C.K. Tittle, Berkely, Calif. McCutchon.

JAEGER, R.M. A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the annual meeting of the N.C.A.R.E. Chapel Hill.

KATZ, D. (1960). The functional approach to the study of attitudes. Publ. Opin. Quart.

KEYS, H. (1934). The influence of Learning and Refution of Weekly as opposed to Monthly Tests. Journal of Educational Psychiatry, 25: 427-436.

KESSLER et al. (1973) Teachers' attitudes towards school psychologists, Psychological Abstracts.

KIM, J. & MUELLER, C.W. (1978) Introduction to Factor Analysis. Beverly Hills: Sage.

KRECH, D., CRUTCHFIELD, R.S. & BALLACHEY, E. (1962). Individual in Society. New York: McGraw-Hill.

LAUSE, A., LEHMANN, I.J., MEHRENS, W.A. (1976) Using Item Analysis to Improve Tests. In Readings in Measurement and Evaluation in Education and Psychology. W.A. Mehrens (ed.), New York: Holt Rinehart and Winston.

LIKERT, R.A. (1932) A technique for the measurement of attitudes. Archives of Psychology, no. 140.

MILLMAN, J. (1974) Criterion-Referenced Measurement. In: Evaluation in Education: Current Applications (ed.) W.J. Popham, Berkeley, Calif.: Mc.Chohan

ROUINELLI, R.J. & HAMBLETON, R.K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. Dutch Journal of Educational Research 2:49-60.

THRUSTONE, K.L. & CHAVE, E.J. (1929). The measurement attitude. Chicago. University of Chicago Press.